IV Всероссийская (национальная) научная конференция с международным участием: «Наука, технологии, общество: Экологический инжиниринг в интересах устойчивого развития территорий» (НТО-IV-2023) при поддержке Краевого фонда науки

10 (2023)

УДК 621-039-542

EDN NJCBAR

Использование нейронных сетей глубокого обучения для классификации токсичных комментариев в социальных сетях

Д.В. Захаренко

Сибирский федеральный университет, Институт космических и информационных технологий, ул. Академика Киренского, 26Б, Красноярск, 660074, Россия

*E-mail: zaharenkodanil1@gmail.com

Аннотация. Целью этого исследования было изучение использования искусственных нейронных сетей глубокого обучения для классификации токсичных комментариев в социальных. Распространенность токсичных взаимодействий на этих платформах достигла небывало высокого уровня, что привело к снижению уровня цифровой цивилизованности. Модераторы этих платформ вынуждены тратить большое количество времени и сил, чтобы контролировать негатив в комментариях. В исследовании рассматриваются различные алгоритмы и методы построения искусственных нейронных сетей, а также сравнивается производительность трех выбранных моделей, чтобы определить наиболее эффективную для решения этой задачи. Классификация характера комментариев, содержащих ненависть, обеспечит платформам гибкость в работе с ними и откроет двери для новых обсуждений и решений.

Ключевые слова: искусственные нейронные сети, глубокое обучение, классификация текста, предобработка текста, токсичные комментарии, социальные сети.

Using deep learning neural networks to classify toxic comments on social media

D.V. Zakharenko

Siberian Federal University, Institute of Space and Information Technologies, ul Akademika Kirenskogo, 26B, Krasnoyarsk, 660074, Russia

*E-mail: zaharenkodanil1@gmail.com

Abstract. The purpose of this study was to study the use of artificial neural networks of deep learning to classify toxic comments on social networks. The prevalence of toxic interactions on these platforms has reached an all-time high level, which has led to a decrease in the level of digital civility. Moderators of these platforms have to spend a lot of time and effort to control the negative in the comments. The study examines various algorithms and methods for building artificial neural networks, and compares the performance of the three selected models to determine the most effective for solving this problem. Classifying the nature of hate comments will provide platforms with flexibility in dealing with them and open the door to new discussions and solutions.

Keywords: artificial neural networks, deep learning, text classification, text preprocessing, toxic comments, social networks.

1. Введение

В наше время большое количество людей общаются друг с другом в комментариях к публикациям в социальных сетях. Возможность комментировать позволяет людям открыто выражать свои мысли и чувства, а также облегчает обмен информацией и помогает людям в поиске решений различных проблем.

Согласно индексу цифровой цивилизованности Microsoft, вежливость онлайн-общения достигла четырехлетнего минимума в 2019 году [1]. Токсичные взаимодействия происходят на платформах социальных сетей, где преобладают жестокие политические дебаты и нападки на личностные качества. То, что должно было быть хорошей социальной силой, платформой для свободного обсуждения в Интернете, превратилось в трясину ненависти. В этом контексте машинное обучение может оказать огромную помощь, поскольку то, на что у человека могут уйти часы работы, может быть выполнено искусственным интеллектом за несколько секунд.

2. Постановка задачи

Исследование включает в себя обзор методов, используемых для достижения целевых результатов, с использованием Python и его библиотек. Оно также охватывает технические аспекты, такие как процесс построения, обучения и оценки моделей искусственных нейронных сетей. Классификация характера комментариев, содержащих ненависть, обеспечит платформам гибкость в работе с ними и откроет двери для новых обсуждений и решений.

В статье будут рассмотрены различные алгоритмы и методы построения искусственных нейронных сетей, а также сравнена производительность трех выбранных моделей, чтобы определить наиболее эффективную для решения этой задачи. Комментарии со страницы обсуждения в Википедии выполняют роль данных для построения моделей классификации.

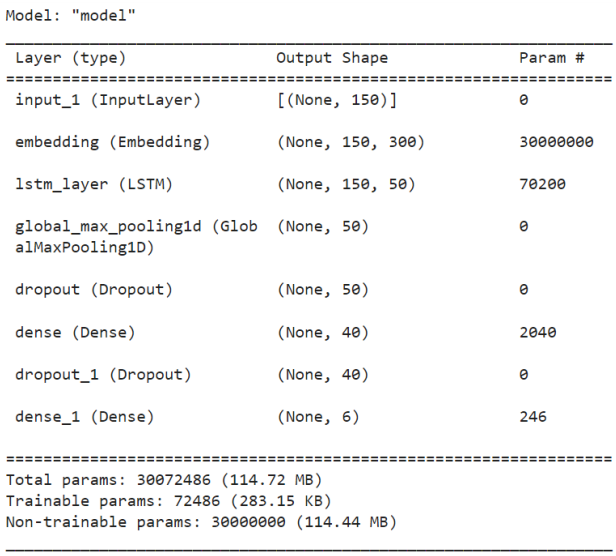
3. Методы и материалы исследования

В этом разделе рассматриваются 3 модели, созданные для обучения, и приводится соответствующая важная информация о различных параметрах, которые можно учитывать при создании модели машинного обучения.

Все 3 модели имеют разные архитектуры. Первая – модель LSTM с 8 слоями, вторая – модель CNN с 8 слоями, третья – гибридная модель LSTM-CNN с 12 слоями.

3.1. LSTM

Первой реализованной моделью стала искусственная нейронная сеть с долгой краткосрочной памятью, поскольку она наиболее согласуется с текстовыми данными. Иерархическое отображение модели LSTM и информация о слоях и нейронах отображены на рисунке 1.



```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 150)]             0

 embedding (Embedding)       (None, 150, 300)          30000000

 lstm_layer (LSTM)           (None, 150, 50)           70200

 global_max_pooling1d (Glob  (None, 50)                0
 alMaxPooling1D)

 dropout (Dropout)           (None, 50)                0

 dense (Dense)               (None, 40)                2040

 dropout_1 (Dropout)         (None, 40)                0

 dense_1 (Dense)             (None, 6)                 246

=================================================================
Total params: 30072486 (114.72 MB)
Trainable params: 72486 (283.15 KB)
Non-trainable params: 30000000 (114.44 MB)
_________________________________________________________________
```

Рисунок 1. Архитектура модели LSTM.

Модель состоит из одного входного слоя, одного embedding слоя, за которым следует один слой LSTM с 50 нейронами. Модель имеет 2 слоя дропаут, 2 полносвязных слоя и один слой подвыборки в порядке, показанном на рисунке 1. Общее количество параметров составило 30 072 486, из которых 72 486 были поддающимися обучению.

Дропаут слой (dropout) – это метод регуляризации, используемый в нейронных сетях для предотвращения переобучения. Во время обучения некоторые нейроны в слоях отсева случайным образом устанавливаются равными нулю, предотвращая чрезмерную зависимость модели от какой-либо одной функции или набора функций. Это способствует повышению эффективности обобщения модели [2].

Полносвязный слой – это основной строительный элемент нейронной сети. Каждый нейрон в одном слое соединен с каждым нейроном в следующем. Полносвязные

слои могут быть многослойными для построения глубоких нейронных сетей и используются для изучения сложных нелинейных корреляций между входными и выходными данными [3].

3.2. CNN

Второй созданной моделью была одномерная сверточная нейронная сеть. Иерархическое отображение модели CNN изображено на рисунке 2.

```
Model: "model_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_2 (InputLayer)        [(None, 150)]             0

 embedding_1 (Embedding)     (None, 150, 300)          30000000

 conv1d (Conv1D)             (None, 146, 128)          192128

 max_pooling1d (MaxPooling1  (None, 29, 128)           0
 D)

 flatten (Flatten)           (None, 3712)              0

 dropout_2 (Dropout)         (None, 3712)              0

 dense_2 (Dense)             (None, 128)               475264

 dense_3 (Dense)             (None, 6)                 774

=================================================================
Total params: 30668166 (116.99 MB)
Trainable params: 668166 (2.55 MB)
Non-trainable params: 30000000 (114.44 MB)
_________________________________________________________________
```

Рисунок 2. Архитектура модели CNN.

Модель состоит из одного входного слоя и одного embedding слоя, за которым следует один слой свертки. Модель имеет 1 дропаут слой, 2 полносвязных слоя, 1 слой подвыборки и один слой flatten в порядке, показанном на рисунке 2. Общее количество параметров составило 30 668 166, из которых 668 166 были поддающимися обучению.

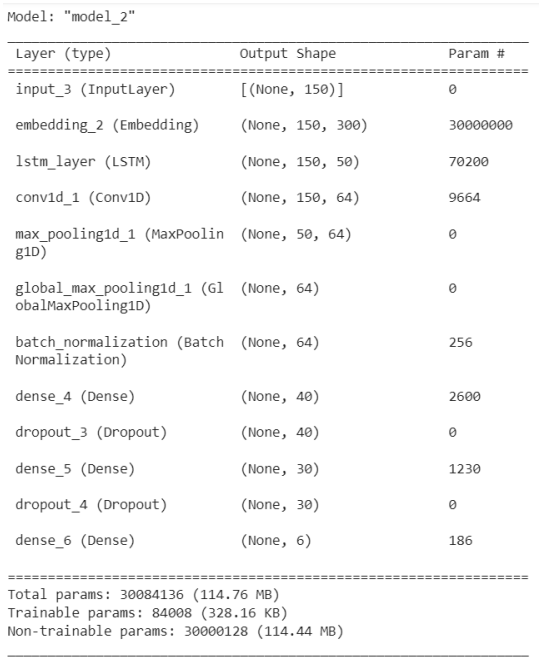
Сверточные нейронные сети (CNN) – это тип нейронных сетей, который доказал свою исключительную эффективность в приложениях распознавания изображений. Однако он был успешно применен в приложениях для обработки естественного языка (NLP), таких как классификация текста. CNN отлично подходит для категоризации текста, поскольку он может распознавать основные свойства текста, такие как n-граммы и фразы, не требуя сложной разработки функций.

Архитектура CNN состоит из последовательности сверточных слоев, которые сканируют входную матрицу на наличие шаблонов или объектов с помощью фильтров. Фильтры применяются к каждому возможному окну ввода слов, позволяя сети изучать

шаблоны в различных масштабах. Выходные данные сверточного слоя впоследствии передаются через слой подвыборки, который уменьшает размерность и собирает наиболее значимую информацию. Наконец, прежде чем принять окончательное решение о классификации, выходные данные слоя подвыборки передаются через один или несколько полносвязных слоев [4].

3.1. LSTM-CNN

Третьей созданной моделью был гибрид долговременной краткосрочной памяти и одномерной сверточной нейронной сети. Иерархическое отображение модели LSTM-CNN и информацию о слоях и нейронах можно увидеть на рисунке 3.



```
Model: "model_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_3 (InputLayer)        [(None, 150)]             0

 embedding_2 (Embedding)     (None, 150, 300)          30000000

 lstm_layer (LSTM)           (None, 150, 50)           70200

 conv1d_1 (Conv1D)           (None, 150, 64)           9664

 max_pooling1d_1 (MaxPoolin  (None, 50, 64)            0
 g1D)

 global_max_pooling1d_1 (Gl  (None, 64)                0
 obalMaxPooling1D)

 batch_normalization (Batch  (None, 64)                256
 Normalization)

 dense_4 (Dense)             (None, 40)                2600

 dropout_3 (Dropout)         (None, 40)                0

 dense_5 (Dense)             (None, 30)                1230

 dropout_4 (Dropout)         (None, 30)                0

 dense_6 (Dense)             (None, 6)                 186

=================================================================
Total params: 30084136 (114.76 MB)
Trainable params: 84008 (328.16 KB)
Non-trainable params: 30000128 (114.44 MB)
_________________________________________________________________
```

Рисунок 3. Архитектура модели LSTM-CNN.

Модель состоит из одного входного слоя, одного embedding слоя, за которым следует один слой LSTM с 50 нейронами и один слой свертки. Модель имеет 2 слоя дропаут, 3 полносвязных слоя, 2 слоя подвыборки и один слой пакетной нормализации в порядке, показанном на рисунке 3. Общее количество параметров составило 30 084 136, из которых 84 008 были поддающимися обучению.

4. Полученные результаты

Для модели LSTM были достигнуты следующие результаты. Потери при обучении 0.0478, потери при валидации 0.0461, точность при обучения 0.9922, точность при валидации 0.9926. Рисунки 4-5 отображают изменение величин точности и потерь во время обучения модели.

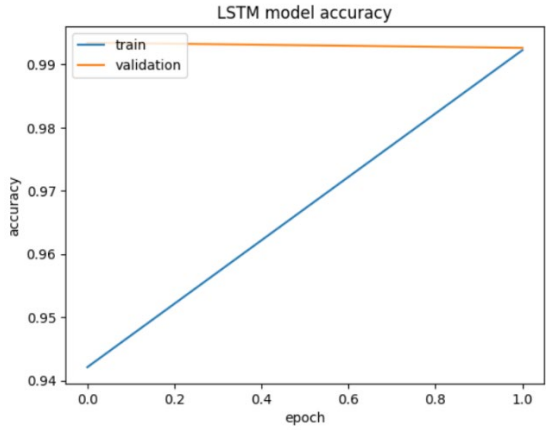


Рисунок 4. Точность модели LSTM.

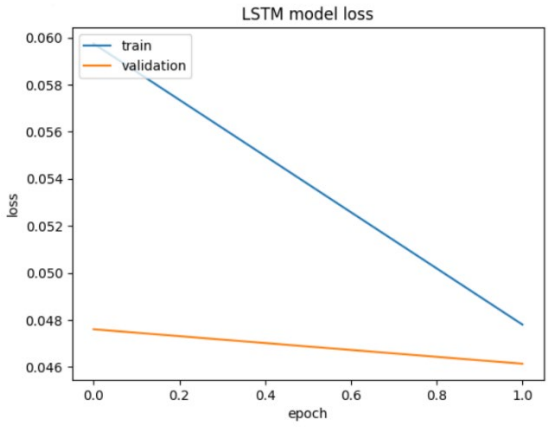


Рисунок 5. Потери модели LSTM.

Оценка точности LSTM модели на публичном тестовом наборе данных равна 0.97327, а на приватном 0.97402. (рисунок 6)

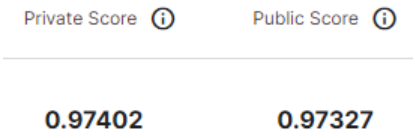


Рисунок 6. Точность модели LSTM на тестовом наборе данных.

Для модели CNN были достигнуты следующие результаты. Потери при обучении 0.0550, потери при валидации 0.0566, точность при обучении 0.9664, точность при валидации 0.9891. Рисунки 7-8 отображают изменение величин точности и потерь во время обучения модели.

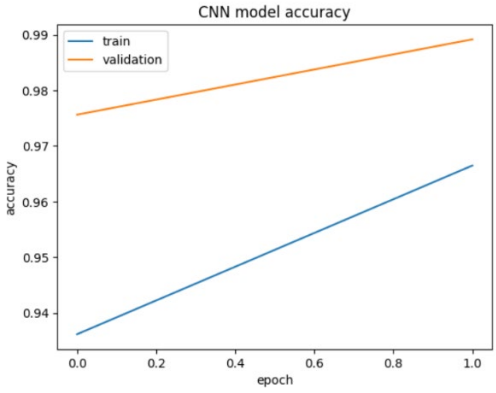


Рисунок 7. Точность модели CNN.

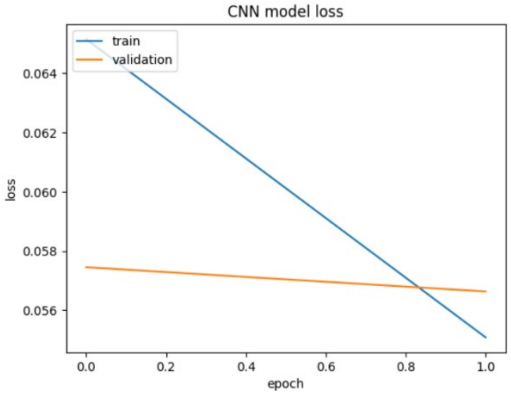


Рисунок 8. Потери модели CNN.

Оценка точности CNN модели на публичном тестовом наборе данных равна 0.95133, а на приватном 0.95273. (рисунок 9)

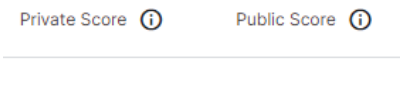


Рисунок 9. Точность модели CNN на тестовом наборе данных.

Для модели LSTM-CNN были достигнуты следующие результаты. Потери при обучении 0.0490, потери при валидации 0.0469, точность при обучения 0.9898, точность при валидации 0.9932. Рисунки 10-11 отображают изменение величин точности и потерь во время обучения модели.

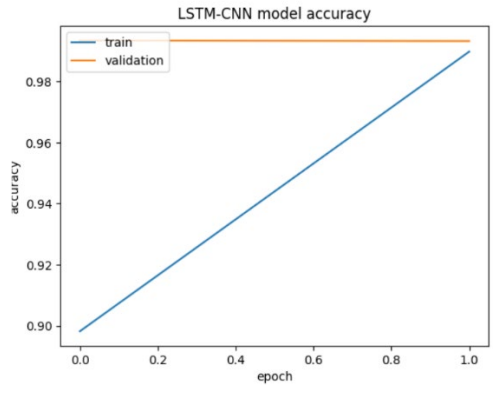


Рисунок 10. Точность модели LSTM-CNN.

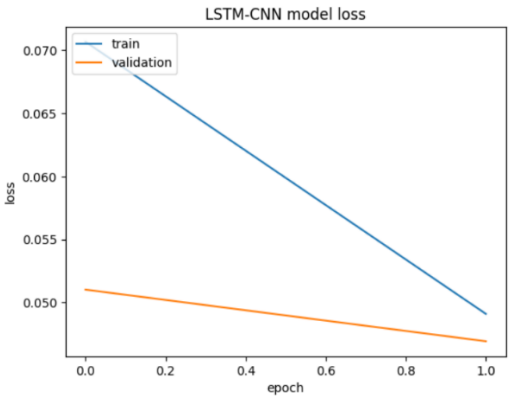


Рисунок 11. Потери модели LSTM-CNN.

Оценка точности LSTM-CNN модели на публичном тестовом наборе данных равна 0.97851, а на приватном 0.97648. (рисунок 12)

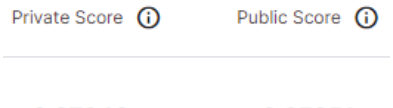


Рисунок 12. Точность модели LSTM-CNN на тестовом наборе данных.

5. Выводы

В результате статьи был выполнен обзор существующих исследований, который показал, что сверточные нейронные сети, рекуррентные нейронные сети и их гибрид являются одними из наиболее эффективных моделей для обнаружения и классификации токсичных комментариев. Эти модели продемонстрировали высокую точность, высокую скорость обработки и надежную производительность в различных контекстах.

Однако использование глубокого обучения и искусственных нейронных сетей для обнаружения токсичных комментариев также сопряжено с некоторыми трудностями. Например, модели необходимо обучать на разнообразных и репрезентативных наборах данных, чтобы гарантировать то, что они могут эффективно обобщаться на новые данные.

Будущие исследования должны быть сосредоточены на решении этих проблем и повышении производительности глубокого обучения и искусственных нейронных сетей для обнаружения токсичных комментариев. Они могут включать в себя изучение новых архитектур и оптимизацию гиперпараметров.

Выводы этой статьи предполагают, что глубокое обучение и искусственные нейронные сети обладают большим потенциалом для решения проблемы токсичных комментариев в социальных сетях. Предоставляя точные и эффективные решения для обнаружения и классификации токсичных комментариев, эти модели могут способствовать здоровому и уважительному взаимодействию в Интернете.

Список литературы

1. Блог Microsoft. Новое исследование Microsoft показывает, что цифровая цивилизованность достигла самого низкого уровня за последние 4 года. – URL

https://blogs.microsoft.com/on-the-issues/2020/02/10/digital-civility-lowest. (дата обращения: 14.09.2023).

2. TensorFlow. tf.keras.layers.Dropout | TensorFlow Core v2.14.0. – URL: https://www.tensorflow.org/api_docs/python/tf/keras/layers/Dropout. (дата обращения: 29.09.2023).
3. TensorFlow. tf.keras.layers.Dense | TensorFlow Core v2.14.0. – URL: https://www.tensorflow.org/api_docs/python/tf/keras/layers/Dense. (дата обращения: 03.10.2023).
4. Randolph. Глубокое обучение для классификации текста с несколькими метками. URL: https://github.com/RandolphVI/Multi-Label-Text-Classification. (дата обращения: 14.10.2023).