Выбор модели для решения задачи автоматической классификации речевой агрессии

И. Е. Воронина, e-mail: irina.voronina@gmail.com
М. К. Пастревич, e-mail: mirstat@mail.ru

Воронежский государственный университет

Аннотация. *В данной работе сравниваются две модели Cointegrated/rubert-tiny2 и SkolkovoInstitute/russian_toxicity_classifier для последующего решения задачи классификации вербальной агрессии.*

Ключевые слова: *классификация речевой агрессии, нейронная сеть, модели.*

Введение

В современном мире существуют различные способы коммуникации, одним из которых является виртуальное общение. С расширением сферы мультимедиа вопросы коммуникативного поведения являются актуальными.

В интернет-пространстве наиболее часто и ярко проявляется вербальная агрессия, преимущественно в комментариях под постами в социальных сетях (например, *ВКонтакте*, *Одноклассники*, *Пикабу*, а также в запрещенных на данный момент в РФ *Facebook, Instagram),* отправляемых сообщениях в мессенджерах *(Whatsapp, Telegram, Viber).* Это связано, в частности с тем, что социальные сети предоставляют пользователям практически полную свободу самовыражения.

Одним из подходов к решению задачи автоматической классификации может стать создание и обучение нейронной сети для автоматического поиска и классификации вербальной агрессии в интернет-пространстве.

Так как русский язык обладает огромным словообразовательным потенциалом и лексико-семантическим разнообразием, это существенно

осложняет решение задачи автоматической классификации вербальной агрессии. Кроме того, при машинном обучении моделей можно столкнуться с проблемой нехватки вычислительных мощностей. Частично с этим справляется персональная версия GPU, а также сервисы, предоставляющие облачные вычисления. Кроме того, существует широкий выбор предобученных на русском языке моделей.

В современной лингвистике существует большое количество классификаций речевой агрессии. В представленном исследовании была взяты следующие виды классификаторов агрессии [1]:

– Эксплетивная. К ней относятся брань, призывы, речевые угрозы, например: «*Феерические идиоты*», «*Дэбилы рагулячие*??».

– Манипулятивная. К ней относят запрет на речь, например: «*закрой рот*».

– Имплицитная. Такой вид речевой агрессии характеризует, например, косвенные речевые акты, иронические инвективы, например: «*Ступай, уже, хватит блистать «интеллектом»*.

Представляется целесообразным добавить нейтральную лексику с перспективой последующего расширения полей классификаторов.

Целью работы является у улучшения точности определения агрессии в текстах комментариев постов в социальных сетях и новостных сайтах.

Методы исследования

Для реализации классификации решено было использовать следующий вид корпуса данных:

$$(x_i, y_i)_i^L = 0 \quad (1)$$

где $x_i \in \mathbb{R}^n$ – i - й комментарий пользователя, а $y_i \in \{1, 2, 3, 4\}$ – метки классов. Будем использовать функцию F, которая для каждого комментария будет ставить соответствующую метку

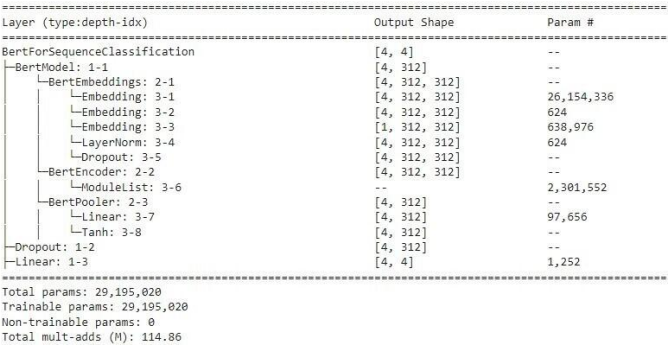
$$F(x_i) = y_i \quad (2)$$

При подготовке корпуса данных был произведен ряд действий:

– обязательное удаление неинформативных символов с помощью стандартной библиотеки Python 3.8 “re”;

– токенизация, лемматизация и приведение к размеру 312 слов.

Архитектура модели состоит из трех основных слоев: BertModel, Dropout, Linear (рис 1).



```
==========================================================================================
Layer (type:depth-idx)                                  Output Shape              Param #
==========================================================================================
BertForSequenceClassification                           [4, 4]                    --
├─BertModel: 1-1                                        [4, 312]                  --
│    └─BertEmbeddings: 2-1                              [4, 312, 312]             --
│    │    └─Embedding: 3-1                              [4, 312, 312]             26,154,336
│    │    └─Embedding: 3-2                              [4, 312, 312]             624
│    │    └─Embedding: 3-3                              [1, 312, 312]             638,976
│    │    └─LayerNorm: 3-4                              [4, 312, 312]             624
│    │    └─Dropout: 3-5                                [4, 312, 312]             --
│    └─BertEncoder: 2-2                                 [4, 312, 312]             --
│    │    └─ModuleList: 3-6                             --                        2,301,552
│    └─BertPooler: 2-3                                  [4, 312]                  --
│    │    └─Linear: 3-7                                 [4, 312]                  97,656
│    │    └─Tanh: 3-8                                   [4, 312]                  --
├─Dropout: 1-2                                          [4, 312]                  --
├─Linear: 1-3                                           [4, 4]                    1,252
==========================================================================================
Total params: 29,195,020
Trainable params: 29,195,020
Non-trainable params: 0
Total mult-adds (M): 114.86
```

Рис. 1. Архитектура модели Cointegrated/rubert-tiny2

На входе для создания векторных представлений используется слой BertEmbedding, содержащий следующие параметры: word_embeddings = 83828 (размер словаря); output = 312 (размер embedding'a), длина входной последовательности = 312. Слой Dropout (p = 0.1) предназначен для уменьшения вероятности переобучения сети. Слой Linear определяет, к какому классу относится комментарий.

В процессе работы был собран и вручную классифицирован набор данных на основе комментариев в социальных сетях *ВКонтакте*, *Одноклассники*, *Пикабу* и комментариев из открытых каналов в *Telegram.* Распределение комментариев представлено на рис.2. Корпус состоит из 76767 комментариев, из которых 66486 нейтральных комментариев, 11 манипулятивных, 8018 эксплицитных, 2252 имплицитных.

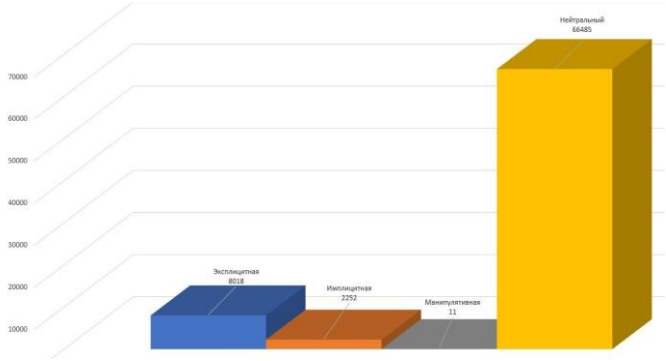


Рис. 2. Распределение комментариев

Были использованы стандартные метрики для оценки качества работы модели:

- Precision (точность).
- Recall (полнота).
- Accuracy (описание точности предсказания модели по всем классам).

Результаты обучения модели Cointegrated/rubert-tiny2 представлены в таблице 1:

Таблица 1

Этапы обучения модели

Количество комментариев	Recall	Precision	Accuracy
11150	0,2541	0,2541	0,7111
26840	0,5432	0,5432	0,7915
39350	0,7567	0,7567	0,8675
56779	0,8541	0,8541	0,9126
76767	0,8462	0,8462	0,911

На рис.3 представлен график потерь в различные моменты обучения модели Cointegrated/rubert-tiny2, иллюстрирующий общее уменьшение при переобучении.

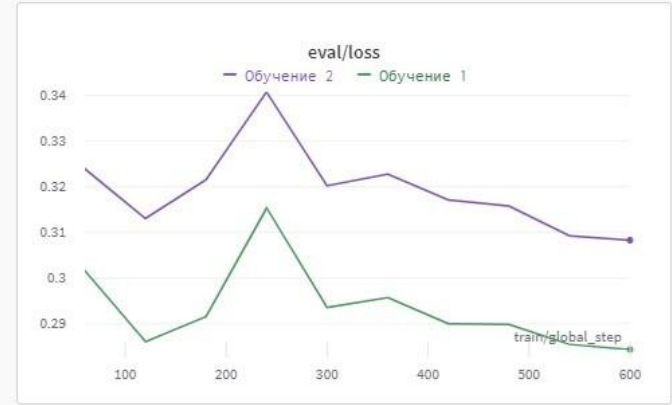


Рис. 3. График потерь при обучении Cointegrated/rubert-tiny2

К сожалению, при каждом новом этапе переобучения точность результата падает.

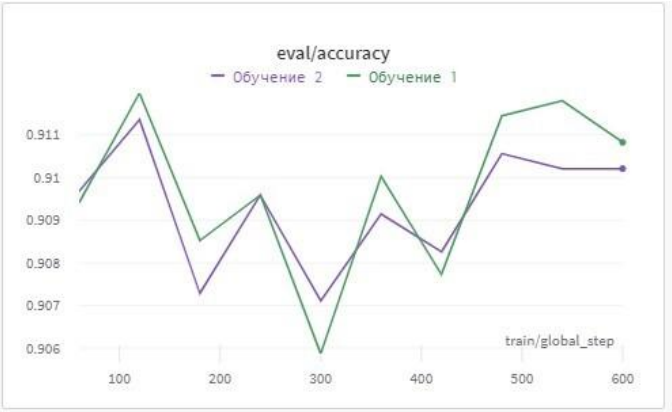
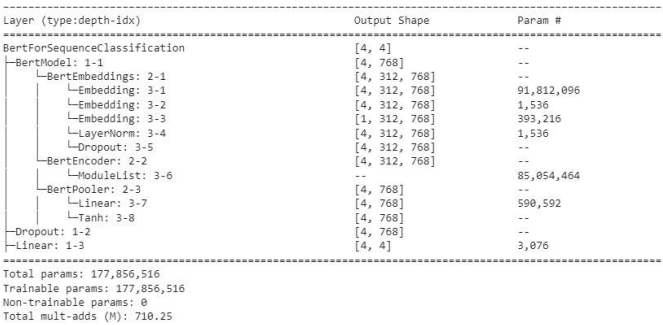


Рис. 4. Отношение точности при обучении Cointegrated/rubert-tiny2

Архитектура модели SkolkovoInstitute/russian_toxicity_classifier имеет также как и Cointegrated/rubert-tiny2 три слоя: BertModel, Dropout, Linear, но обладает иными параметрами (рис.5):



```
==========================================================================================
Layer (type:depth-idx)                                  Output Shape              Param #
==========================================================================================
BertForSequenceClassification                           [4, 4]                    --
├─BertModel: 1-1                                        [4, 768]                  --
│    └─BertEmbeddings: 2-1                              [4, 312, 768]             --
│    │    └─Embedding: 3-1                              [4, 312, 768]             91,812,096
│    │    └─Embedding: 3-2                              [4, 312, 768]             1,536
│    │    └─Embedding: 3-3                              [1, 312, 768]             393,216
│    │    └─LayerNorm: 3-4                              [4, 312, 768]             1,536
│    │    └─Dropout: 3-5                                [4, 312, 768]             --
│    └─BertEncoder: 2-2                                 [4, 312, 768]             --
│    │    └─ModuleList: 3-6                             --                        85,054,464
│    └─BertPooler: 2-3                                  [4, 768]                  --
│    │    └─Linear: 3-7                                 [4, 768]                  590,592
│    │    └─Tanh: 3-8                                   [4, 768]                  --
├─Dropout: 1-2                                          [4, 768]                  --
├─Linear: 1-3                                           [4, 4]                    3,076
==========================================================================================
Total params: 177,856,516
Trainable params: 177,856,516
Non-trainable params: 0
Total mult-adds (M): 710.25
```

Рис. 5. Архитектура модели SkolkovoInstitute/russian_toxicity_classifier

Слой BertEmbedding содержит следующие параметры: word_embeddings = 119547 (размер словаря); output = 768 (размер embedding'a), длина входной последовательности = 768. Слой Dropout (p = 0.1) предназначен для уменьшения вероятности переобучения сети.

При обучении модели на вышеупомянутом наборе данных, получены следующие результаты: Recall = 0,8862; Precision = 0,8862; Accuracy =0,9265.

На рис.6 представлен график потерь в различные моменты обучения модели SkolkovoInstitute/russian_toxicity_classifier. Мы наблюдаем незначительный рост потерь, а затем уменьшение.

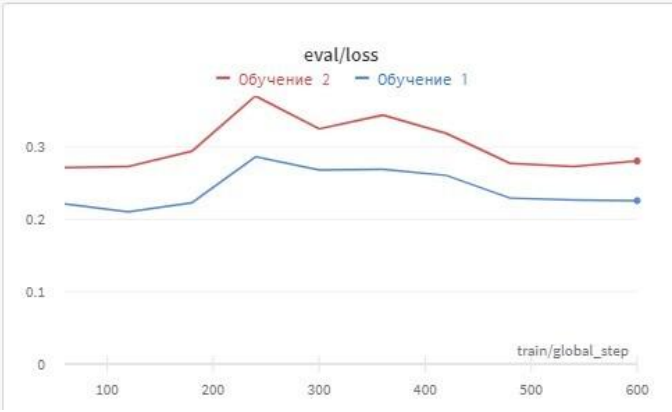


Рис. 6. График потерь при обучении SkolkovoInstitute/russian_toxicity_classifier

На рис.7 представлен график точности при обучении модели SkolkovoInstitute/russian_toxicity_classifier, который демонстрирует при переобучении снижение точности.

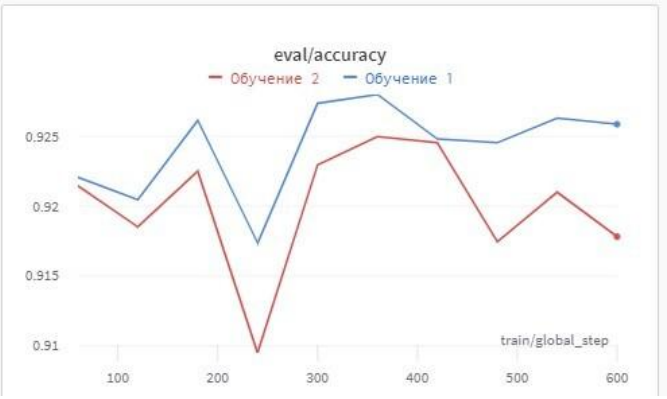


Рис. 7. Отношение точности при обучении SkolkovoInstitute/russian_toxicity_classifier

На рис. 8 предоставлен сравнительный график точности обучения двух моделей BertForSequenceClassification и SkolkovoInstitute/russian_toxicity_classifier.

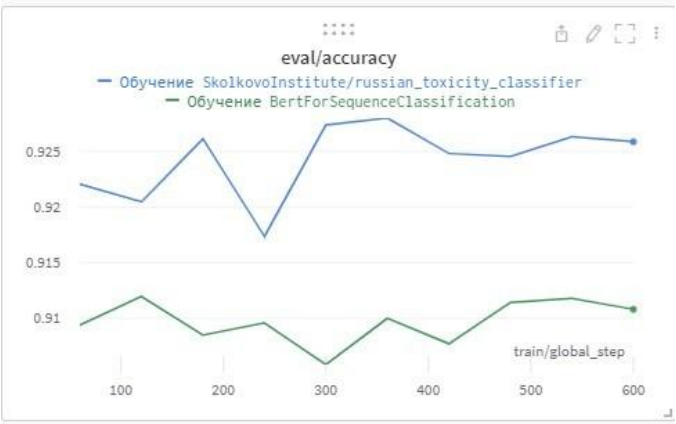


Рис. 8. Сравнительный график точности

Заключение

Как можем заметить, точность обучения в предобученной модели SkolkovoInstitute/russian_toxicity_classifier существенно выше, чем у Cointegrated/rubert-tiny.

По результатам проведенного исследования можно сделать следующий вывод: для решения задачи классификацию вербальной агрессии будет использоваться модель SkolkovoInstitute/russian_toxicity_classifie. Предполагается значительное расширение корпуса данных.

Список литературы

1. Шейгал Е.И. Семиотика политического дискурса: монография / Е.И.Шейгал. - Волгоград: Перемена, 2000. - 367 с.

2. Лыченко Н.М., Сорочинская А.В., Сравнение эффективности методов векторного представления слов для определения тональности текстов [Электронный ресурс] : CyberLeninka. – Режим доступа:

https://cyberleninka.ru/article/n/sravnenie-effektivnosti-metodov-vektornogo-predstavleniya-slov-dlya-opredeleniya-tonalnosti-tekstov/viewer

3. ТАСС. "Одноклассники" запустили нейросеть для борьбы с агрессивными комментариями [Электронный ресурс] : официальный сайт. – Режим доступа : https://tass.ru/ekonomika/13977023

4. Определение токсичных комментариев на русском языке [Электронный ресурс] : официальный сайт. – Режим доступа : https://habr.com/ru/company/vk/blog/526268/

5. Hugging Face [Электронный ресурс] : официальный сайт. – Режим доступа : https://huggingface.co/cointegrated/rubert-tiny2