Выбор модели для решения задачи автоматической классификации речевой агрессии

Воронина И.Е., д.т.н. проф. каф. ПОиАИС ВГУ

Пастревич М.К., асп.

Актуальность работы

- Расширение влияния сферы массмедиа на коммуникативное поведение человека.
- Задача снижения речевой агрессии в информационном пространстве.
- Создание корпуса данных для решения задачи определения агрессивного контента.

Виды агрессии для классификаторов

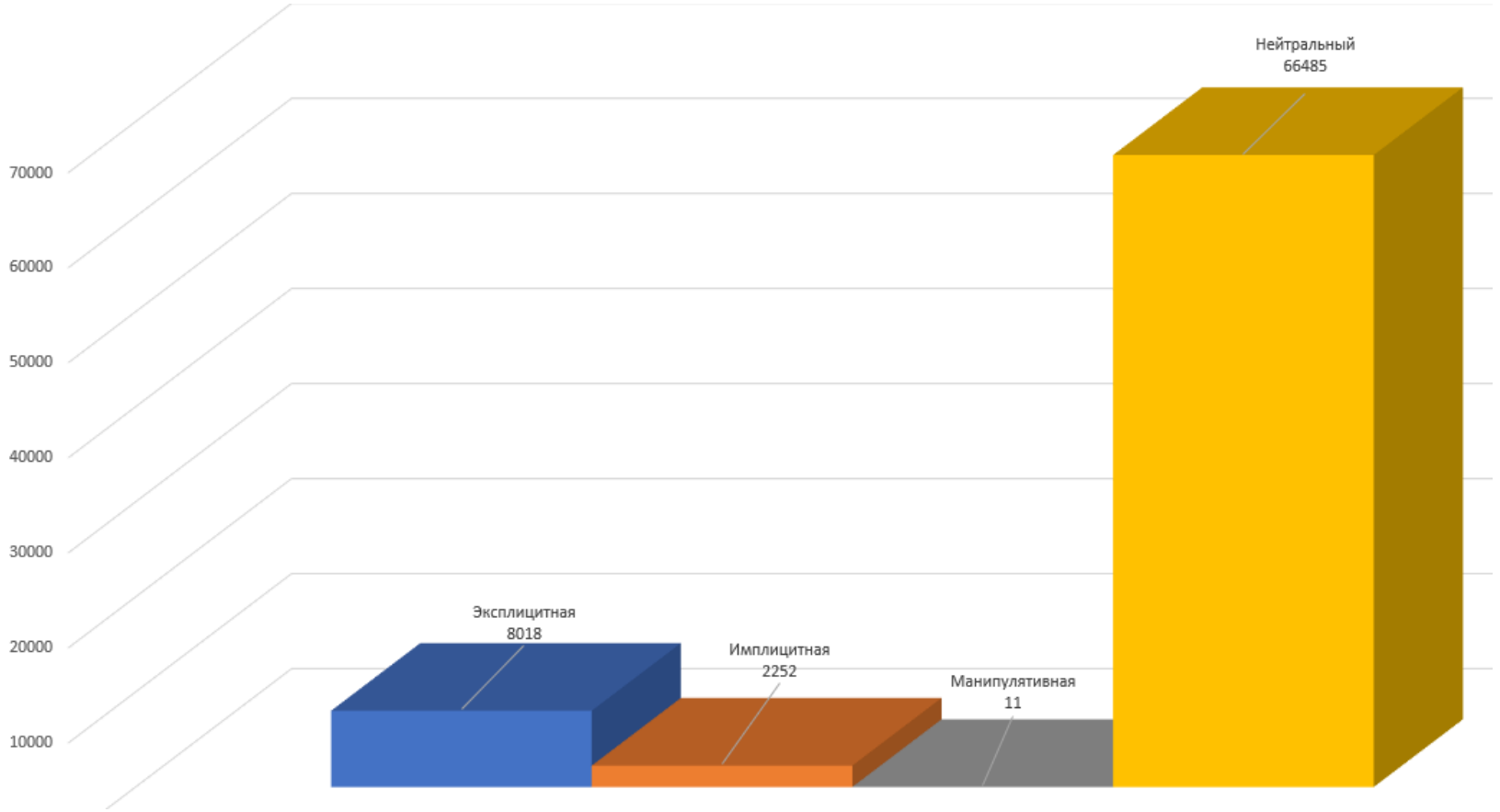
- Эксплетивная: брань, призывы, речевые угрозы, например: «Феерические идиоты», «Дэбилы рагулячие??».
- Манипулятивная: запрет на речь, например: «закрой рот».
- Имплицитная: характеризует, например, косвенные речевые акты, иронические инвективы, например: «Ступай, уже, хватит блистать «интеллектом».

Методы исследования

Вид корпуса данных: =0, где – i-й комментарий пользователя, а {1,2,3,4}- метки классов.

Будем использовать функцию F, которая для каждого комментария будет ставить соответствующую метку)=

Распределение комментариев



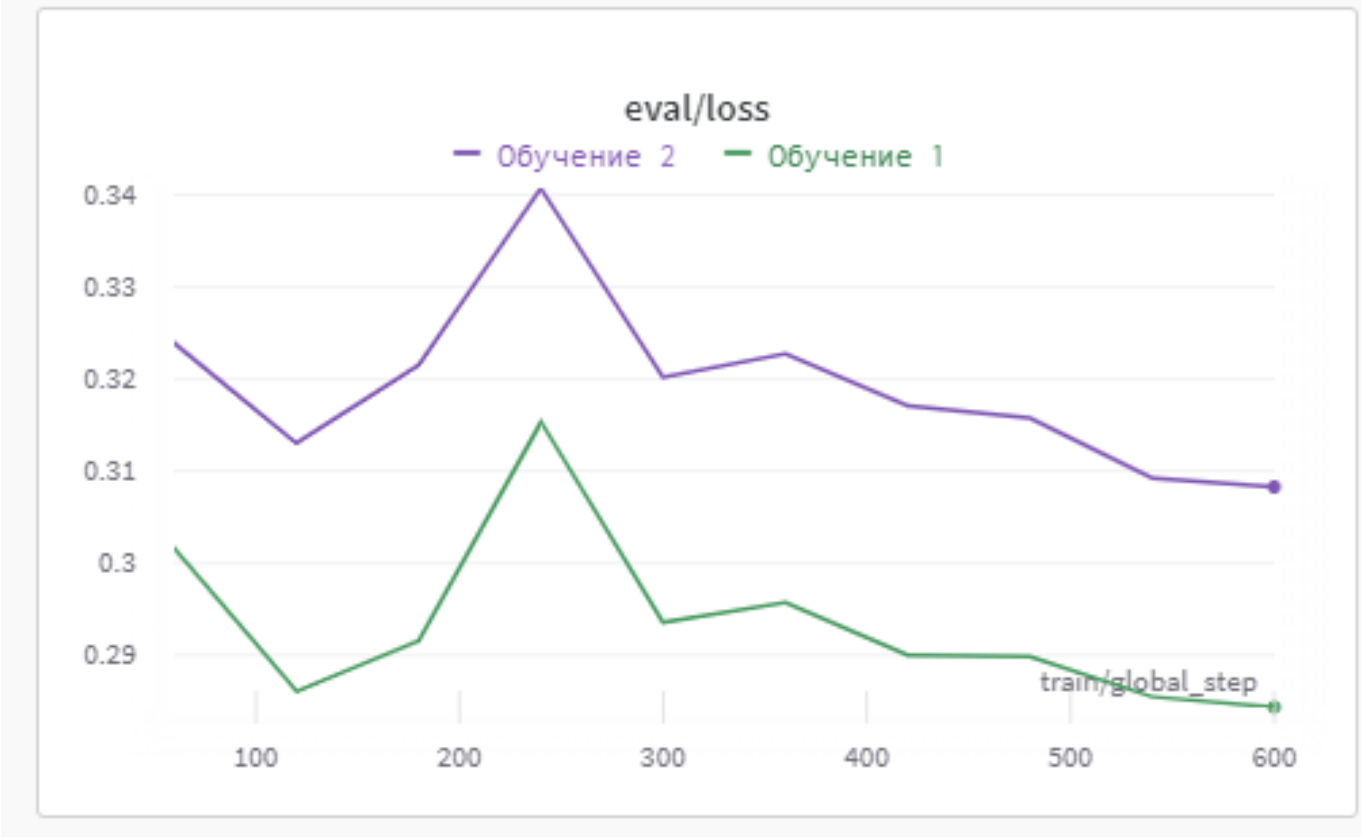
Корпус данных, собранный на основе социальных сетй *ВК, Одноклассники, Пикабу* состоит из 76767 комментариев, из которых 66486 нейтральных, 11 манипулятивных, 8018 эксплицитных, 2252 имплицитных.

Архитектура модели cointegrated/rubert-tiny2

```
==========================================================================================================
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
==========================================================================================================
```

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.

График потерь при обучении Cointegrated/rubert-tiny2



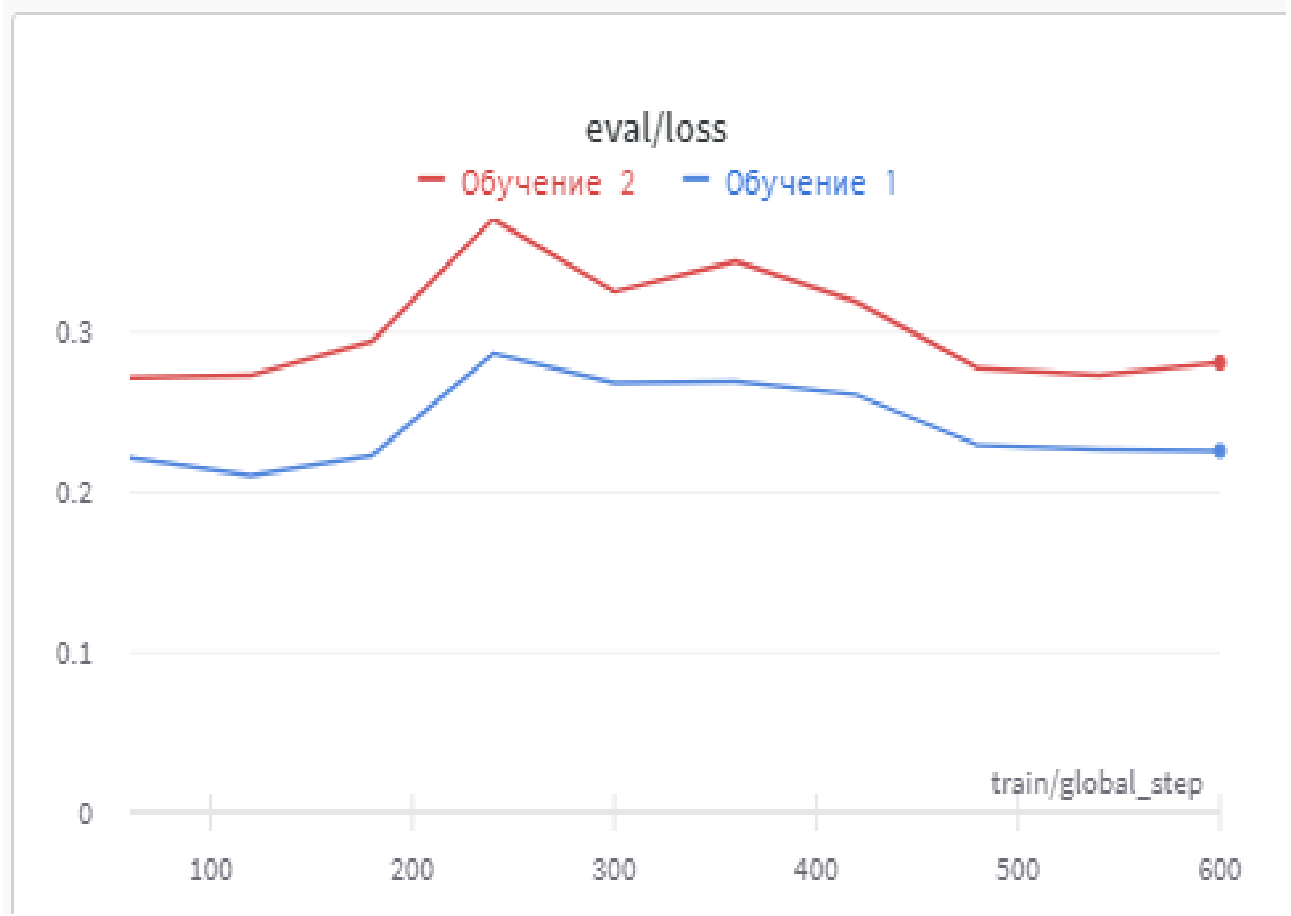
Отношение точности при обучении Cointegrated/rubert-tiny2

Архитектура модели SkolkovoInstitute/russian_toxicity_classifier

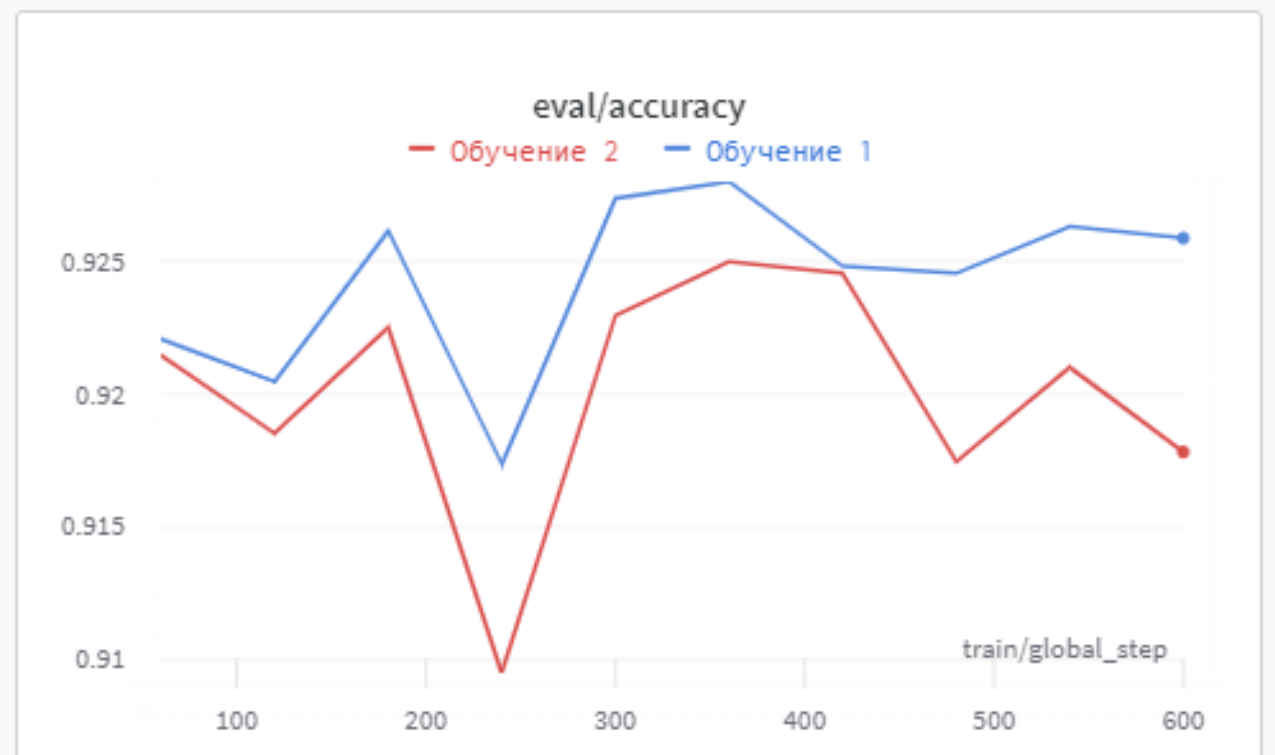
```
==========================================================================================
Layer (type:depth-idx)                                  Output Shape              Param #
==========================================================================================
BertForSequenceClassification                           [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
```

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.

График потерь при обучении SkolkovoInstitute/russian_toxicity_classifier



Отношение точности при обучении SkolkovoInstitute/russian_toxicity_classifier



Сравнительный график точности

Вывод

Для решения задачи классификации вербальной агрессии будет использоваться модель SkolkovoInstitute/russian_toxicity_classifie.

Предполагается значительное расширение корпуса данных.