**DOI 10.36622/VSTU.2022.25.3.003**
**УДК 519.766.4:004.032.26**

# СРАВНЕНИЕ МОДЕЛЕЙ-ТРАНСФОРМЕРОВ BERT ПРИ ВЫЯВЛЕНИИ ДЕСТРУКТИВНОГО КОНТЕНТА В СОЦИАЛЬНЫХ МЕДИА

**В.А. Минаев, А.В. Симонов**

Цель статьи состоит в определении наиболее эффективной модели из семейства BERT по выявлению деструктивного контента в социальных медиа. Произведено сравнение пяти наиболее известных моделей BERT по выявлению деструктивного контента. Для этого осуществлено создание текстового корпуса из материалов социальных медиа (СМ), дополненного запрещённым к распространению в Российской Федерации контентом нацистского характера из Федерального списка экстремистских материалов. Представлена структура классификатора текстовых данных, основанного на глубокой искусственной нейронной сети BERT, и описана его работа на каждом этапе. Проведен поиск наиболее эффективного метода предварительной обработки текстов. Оценена эффективность работы различных голов классификаторов, основанных на трансформере BERT. Оценено влияние дообучения модели BERT и доказана эффективность его применения с расчетом перплексии. Представлены сравнительные таблицы работы классификаторов на каждом этапе исследования. Найдена наиболее эффективная архитектура классификатора на основе трансформера BERT, выполняющего задачу выявления деструктивного контента с точностью 96,99%.

Ключевые слова: деструктивный текстовый контент, экстремизм, социальные медиа, трансформер BERT, точность классификации.

## Введение

В эпоху стремительного развития цифровых технологий многие перенесли свою активность в Интернет сферу. К ней относится бытовое и деловое общение, которое всё чаще происходит социальные медиа (СМ), к которым относятся социальные сети, форумы, а также различные сервисы обмена мгновенными сообщениями (Telegram, WhatsApp, Viber и т. п.).

В свою очередь, СМ используются не только для обычного общения по интересам, но и активно применяются в целях проведении информационных войн, агитации, распространении недостоверной информации (фейков) и пропаганды экстремистских и иных запрещенных законом идей.

В цифро-ориентированном обществе защита от такого использования СМ, а именно, применения деструктивных информационных воздействий является одним из важнейших факторов, способных влиять на психическое здоровье населения и стабильность функционирования государств. Распространение запрещенного контента отмечено в Доктрине информационной безопасности Российской Федерации, утвержденной Указом Президента Российской Федерации № 646 от 05.12.2016, как одна из основных информационных угроз

Наиболее распространенным видом информации является текстовый (за счет легкости создании и передачи). Учитывая высокую интенсивность генерации новых текстов и постов в СМ (более миллиарда в месяц), осуществлять проверку каждого из них вручную невозможно.

Для этого используются автоматизированные системы (АС), которые можно разделить на два вида – использующие:

- поиск по ключевым словам, который прост в настройке и применении, но имеет высокий уровень ложных срабатываний, т.к. ключевые слова могут попадаться и в нейтральных публикациях;
- методы машинного обучения, показывающие более высокую эффективность по сравнению с предыдущими. Однако во многом их качество

- зависит от конкретных алгоритмов, которые используются для распознавания деструктивного контента (ДК).

В работе [1] приводится сравнение эффективных методов глубоких искусственных нейронных сетей (ГИНС) для выявления деструктивного контента. По результатам экспериментов сделан вывод, что наилучшие результаты по его выявлению показала модель BERT, разработанная в 2017 году.

Целью настоящей статьи является экспериментальный поиск наиболее эффективной архитектуры из моделей семейства BERT, используемой при поиске и идентификации ДК нацистской направленности.

## Методы и этапы экспериментальных исследований

Модели выявления ДК в СМ представляют классификаторы BERT, обученные на специальном текстовом корпусе.

Классификатор представляет следующую последовательность обработки данных:

1. Формирование обучающего корпуса текстов.
2. Предварительная обработка текстов.
3. Классификация векторов предложений.
4. Токенизация и преобразование текста в вектор.
5. Доменная адаптация.

Для того, чтобы найти наиболее подходящую архитектуру классификатора, на каждом этапе проведены исследования его реализации, передавая на следующий этап лучший вариант предыдущего этапа.

Дадим описание экспериментов на каждом этапе исследования классификатора.

*Формирование обучающей выборки*.

В связи с отсутствием в открытом доступе выборки реального ДК, распространяемого в СМ, авторами создан специальный корпус текстов, состоящий из следующих частей.

Корпус СМ (СМК), содержащий посты пользователей и новостные тексты, представляет нейтральный контент СМ.

Нацистский корпус (НК) – корпус, содержащий запрещенные к распространению тексты нацистского содержания, находящиеся в Федеральном списке экстремистских материалов (ФСЭМ) [2].

Шумовой корпус (ШК) – корпус, содержащий похожий по частотности использования слов в НК, а также содержащий частое упоминание слов-сущностей – имен собственных, встречающихся в НК, таких как «Гитлер», «фюрер», «НСДАП» и прочих.

Данный корпус используется в обучении в целях снижения ошибок 1-го и 2-го рода. Особенности его применения описаны в работе [3].

Статистические данные об объеме корпусов приведены в табл. 1.

Таблица 1

Объемы текстовых корпусов

| Корпус | Количество предложений |
|---|---|
| СМК | 109258 |
| НК | 108437 |
| ШК | 108524 |

*Предварительная обработка текстов*.

Предобработка текстов заключается в избавлении от шума в их исследуемых корпусах: удалении спецсимволов и шумовых слов (союзов, предлогов), не влияющих на смысл; приведении текста к единому регистру; специальной обработке слов.

Для повышения качества работы классификатора применялись три подхода:

1. Легкая обработка. Приведение всех слов к единому регистру и удаление специальных символов. Слова при этом не изменялись.

2. Легкая обработка с последующим применением стемминга – процедурой замены слова на его основу, которая не всегда совпадает с морфологическим корнем.

3. Легкая обработка с последующим применением лемматизации – процедурой замены слова на его словарную форму.

*Классификация векторов предложений*.

При поиске наиболее эффективной головы классификатора, встраиваемой в исследуемую модель, выбраны три варианта:

1. Однослойная нейронная сеть прямого распространения из библиотеки huggingface;

2. Логистическая регрессия как один из простых и эффективных методов машинного обучения, используемых в задачах классификации;

3. Двунаправленная ГИНС, с длинной цепью элементов краткосрочной памяти (BiLSTM), которая часто используется при классификации связанных между собой последовательностей (сигналы, тексты) [4].

В качестве модели BERT для эксперимента с заменой головы классификатора выбран ruBERT-base.

*Токенизация и преобразование текста в вектор*.

Токенизация текста при использовании моделей трансформеров – это процесс дробления слов на части (обычно в 2-3 символа) с последующей их заменой на заранее определенное числовое значение. После токенизации на вход BERT подается массив чисел, соответствующий предоработанному и токенизированному тексту. Каждая модель BERT имеет свой собственный токенизатор, подходящий только ей самой, поэтому далее, говоря о конкретных моделях BERT, будем иметь ввиду и их токенизатор.

Трансформер BERT – ГИНС, состоящая из нескольких последовательных нейронных сетей-кодировщиков, использующих механизм, который позволяет учитывает контекст соседних токенов в предложении [5].

BERT может использоваться в различных задачах: определение похожести предложений, угадывание пропущенного или следующего слова и др., однако в случае с использованием его в качестве классификатора, после проведения операции токенизации, данные проходят слои кодировщиков, на выходе имея векторное представление предложения, которое подается уже на голову классификатора.

Голова классификатора – составная часть модели BERT, представляющая собой нейронную сеть, которая занимается классификацией поступающих на нее данных. На данном этапе эксперимента в качестве головы выбрана встроенная по умолчанию однослойная нейронная сеть прямого распространения.

В качестве исследуемых моделей BERT в эксперименте использованы модели, приведенные ниже.

RuBERT-base – стандартная модель (12 слоев, 768 размер текстового вектора (эмбеддинга), 180 миллионов параметров), обученная на русской части Википедии и новостных данных [6].

RuBERT-conversational – стандартная модель, обученная на русскоязычном сегменте социальных медиа [7, 8].

SBERT-large-nlu-ru – крупная модель (24 слоя, 1024 размер эмбеддинга, 427 миллионов параметров), обученный на 16 миллиардах русскоязычных токенов, в который вошли как публично доступные данные из СМ и Википедии, так и проприетарные датасеты [9].

Multilingual-BERT – стандартная модель, обученная на корпусе Википедии на 104 языках мира (мультиязычная модель).

Rubert-tiny2 – дистиллированная модель BERT (3 слоя, 312 размер эмбеддинга, 12 миллионов параметров), обученная как на русскоязычных так и на англоязычных корпусах с использованием и ruBERT-base. Дистилляция – способ трансляции обученных знаний из громоздких моделей, на примере стандартных BERT, в меньшие [10].

Все выше названные модели взяты с репозитория huggingface.

*Доменная адаптация*.

Каждая модель трансформера проходила процедуру «первичного» обучения, т.е. нейронная сеть BERT обучалась на больших корпусах текстов определенных языков.

После этого BERT начинает «понимать» семантику и взаимосвязь стоящих рядом токенов в контексте предложений и текстов.

В некоторых случаях для повышения качества работы модели проводят доменную адаптацию – процесс дообучения уже обученных моделей BERT на тематических корпусах текстов. Это необходимо, когда BERT используют в задачах обработки специфических текстов, например, медицинский контент, научно-технические статьи специального профиля со своим специфическим синтаксисом.

В эксперименте проверена гипотеза улучшения качества работы классификатора при дообучении выбранного трансформера BERT с использованием техники Masked Language Modeling (MLM).

Суть MLM заключается в обучении модели правильно предугадывать стоящие рядом токены.

Качество работы классификаторов оценивалось на основе F-меры [11], достоинство которой состоит в том, что она позволяет учитывать точность и полноту и классификации одновременно.

Также оценивалось качество доменной адаптации – насколько правильно модель научилась угадывать замаскированные слова в предложениях. Для этого использована оценка перплексии [12].

Чем ниже значение перплексии, тем лучше осуществляется предсказание токена и, соответственно, лучше работает модель в режиме MLM, а также в качестве классификаторов. Измерять перплексию целесообразно при дообучении моделей и их доменной адаптации.

**Результаты экспериментов**

Результаты эксперимента 1 по определению наиболее эффективного метода предобработки текста представлены в табл. 2, где в качестве классификатора по умолчанию использовался Rubert-base, а в качестве головы – однослойная нейронная сеть прямого распространения.

Таблица 2

Результаты классификации с различными методами предобработки текста

| **Метод предобработки** | ***F*-мера** |
|---|---|
| Лемматизация | 0.9378 |
| Стемминг | 0.9326 |
| Легкая предобработка | **0.9577** |

Из табл. 2 видно – чем больше предобработка меняет реально используемые слова, тем хуже классификация текстов.

Модель BERT учитывает предлоги, окончания, приставки, суффиксы, также извлекая из них полезную информацию, чего нельзя сказать о других ГИНС. Отметим, что методы машинного обучения, такие, например, как логистическая регрессия, классификацию осуществляют более качественно при лемматизации.

Результаты эксперимента 2 по поиску наиболее результативной головы классификатора приведены табл. 3.

Таблица 3

Результаты классификации текстов при различных головах классификатора BERT

| **Голова классификатора** | ***F*-мера** |
|---|---|
| Однослойная нейронная сеть | **0,9577** |
| Логистическая регрессия | 0,9013 |
| BiLSTM | 0,8595 |

Из табл. 3 видно, что лучше всего классифицирует обычная однослойная нейронная сеть прямого распространения, по сравнению с логистической регрессией или BiLSTM.

Низкие результаты BiLSTM может быть связана с тем, что на выходе BERT формируется векторное представление предложения, в котором элементы вектора не связаны между собой. Поэтому в данном случае достоинства BiLSTM как ГИНС с краткосрочной памятью не проявляются в связи.

Результаты эксперимента 3 по поиску наиболее точной модели BERT в качестве классификатора представлены в табл. 4.

В каждом классификаторе использовалась в качестве головы однослойная нейронная сеть, а данные поступали на токенизатор после прохождения процедуры легкой обработки текста.

Таблица 4

Результаты классификации с различными моделями BERT

| **Модель BERT** | ***F*-мера** |
|---|---|
| RuBERT-base | 0, 9577 |
| RuBERT-conversational | 0,9548 |
| SBERT-large-nlu-ru | 0,9367 |
| Multilingual-BERT | 0,9182 |
| Rubert-tiny2 | **0,96** |

Итак, наилучший результат показала дистиллированная модель Rubert-tiny2, обойдя модель базовую модель RuBERT-base, RuBERT-conversational, SBERT-large-nlu-ru и многоязыковую модель Multilingual-BERT. Стоит также отметить, что Rubert-tiny2 “легче”, чем другие стандартные модели, и обучается в 4 раза быстрее.

В эксперименте 5 определено влияние доменной адаптации на качество распознавания ДК. Сравним при этом близкие по результатам модели Rubert-tiny2 и RuBERT-base.

Обучены модели на том же текстовом корпусе из ФСЭМ [***], который использован в процессе классификации. Результаты эксперимента с расчетом перплексии до и после применения доменной адаптации представлены в табл. 5.

Как мы видим $PP$ изначально ниже у ruBERT-base – это означает, что данная модель должна лучше рассчитывать вектора слов и предложений, в отличие от ruBERT-tiny2, т.е. быть более «семантически развитой» по отношению к обработке естественного языка.

Таблица 5

Результаты применения доменной адаптации моделей ruBERT-base и ruBERT-tiny2

| **Модель BERT** | **Перплексия до адаптации** | **Перплексия после адаптации** |
|---|---|---|
| ruBERT-base | 23,79 | 9,57 |
| ruBERT-tiny2 | 62,06 | 36,89 |

После обучения перплексия снизилась в модели ruBERT-base в 2,5 раза, а в модели ruBERT-tiny2 – в 1,7 раза. Но при этом у второй модели перплексия всё равно значительно выше, чем у первой модели.

После проведения процедуры доменной адаптации с выбранными моделями BERT оценим влияние дообучения на качество классификации. Результаты приведены в табл. 6.

Таблица 6

Результаты классификации по моделям RuBERT-base и RuBERT-tiny2 после процедуры доменной адаптации

| **Модель BERT** | ***F*-мера** |
|---|---|
| RuBERT-base | **0,9699** |
| Rubert-tiny2 | 0,9609 |

## Обсуждение и выводы

Как видно из таблицы 6 доменная адаптация позволила повысить *F*-меру классификации модели ruBERT-base на 1,22%, однако для модели Rubert-tiny2 доменная адаптация не принесла какого-либо значительного роста результатов.

В статье приведены результаты экспериментов по определению наиболее эффективного классификатора на каждом этапе обработки текста, предназначенного для выявления деструктивного контента.

Для достижения данной цели создан корпус из текстов СМ, в который внедрен деструктивный контент нацистского содержания из материалов, входящих в ФСЭМ [2].

Исследованы наиболее популярные методы предобработки текста, головы классификаторов в моделях BERT, а также сами модели BERT, созданные разными командами разработчиков. Проведен эксперимент по применению доменной адаптацией и оценки её влияние на качество работы классификаторов.

Эксперименты показали, что при решении задачи классификации с использованием моделей-трансформеров BERT, наилучшей на этапе предобработки текстов легкая обработка, которая не видоизменяет слова, приводя их к единому регистру.

На этапе токенизации и преобразования текста в вектор лучшей является RuBERT-tiny2 при отсутствии доменной адаптации и RuBERT-base при использовании доменной адаптации.

Наиболее результативной головой на этапе классификации текста является однослойная однонаправленная нейронная сеть. Доменная адаптация может повысить качество работы классификатора, если используются стандартные модели BERT.

Так, итоговая *F*-мера для модели RuBERT-base составила весьма высокий показатель, равный 96,99%.

Полученные результаты по оценке работы моделей выявления ДК целесообразно учитывать при проектировании и конструировании наполнении систем мониторинга и выявления деструктивной информации в СМ.

Дальнейшие перспективы развития и применения модели BERT направлены на разработку:

- универсальной модели по выявлению деструктивного контента различных категорий: радикальный ислам, реабилитация нацизма, античеловеческие культы и секты, антисемитизм и др.;
- модели, позволяющей оценивать деструктивность текстов любых объемов, которые могут в значительной мере превышать по объему максимальный токенизированный вектор BERT.

## Список литературы

1. Минаев В.А., Поликарпов Е.С., Симонов А.В. Применение глубинных нейронных сетей для выявления деструктивного контента в социальных медиа // Информация и безопасность. 2021. Т. 24. № 3. С. 361-372.

2. Экстремистские материалы: Министерство юстиции Российской Федерации. URL: https://minjust.gov.ru/ru/extremist-materials/ (дата обращения: 09.08.2022).

3. Минаев В.А., Симонов А.В. Повышение точности идентификации контента экстремистского характера в социальных медиа // Материалы V Международной научно-практической конференции «Информационная безопасность: вчера, сегодня, завтра». 14 апреля 2022 года. М.: Российский государственный гуманитарный университет, 2022. С. 80-86.

4. Chen T. et al. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN // Expert Systems with Applications. – 2017. – Т. 72. – P. 221-230.

5. Devlin J. et al. Bert: Pretraining of Beep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv: 1810. 04805. 2018. P. 1-16.

6. Kuratov Y., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // arXiv preprint arXiv:1905.07213. 2019. P.1-7.

7. Lison P., Tiedemann J. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles // 10th Conference on International Language Resources and Evaluation (LREC'16). – European Language Resources Association, 2016. – P. 923-929.

8. Shavrina T., Shapovalova O. To the Methodology of Corpus Construct Tree Corpus and Parser // Proceedings of the “Corpora. 2017. P. 78-84.

9. Обучение модели естественного языка с BERT и Tensor Flow // URL: https://habr.com/ru/company/sberdevices/blog/527576/ (дата обращения: 09.08.2022).

10. Hinton G. et al. Distilling the Knowledge in a Neural Network //arXiv preprint arXiv:1503.02531. 2015. Т. 2. №. 7. P.1-9.

11. Sasaki Y. et al. The Truth of the F-measure //Teach Tutor Mater. 2007. Т. 1. №. 5. P. 1-5.

12. Asuncion A., Welling M., Smyth P., Teh Y. W. On Smoothing and Inference for Topic Models //Proceedings of the Twenty-fifth Conference on Uncertainty in Artificial Intelligence // AUAI Press. 2009. P. 27-34.

Московский университет МВД России им. В.Я. Кикотя
V. Ya. Kikot Moscow University of the Russian Internal Affairs Ministry

Московский государственный технический университет им. Н. Э. Баумана
Bauman Moscow State Technical University

Поступила в редакцию 23.08.2022

**Информация об авторах**

**Минаев Владимир Александрович** – д-р техн. наук, профессор, профессор кафедры специальных информационных технологий, Московский университет МВД РФ им. В.Я. Кикотя, e-mail: m1va@yandex.ru
**Симонов Александр Валерьевич** – аспирант кафедры защиты информации, Московский государственный технический университет им. Н. Э. Баумана, e-mail: san.siman@yandex.ru

## COMPARISON OF BERT TRANSFORMER MODELS IN IDENTIFYING DESTRUCTIVE CONTENT IN SOCIAL MEDIA

**V.A. Minaev, A.V. Simonov**

The purpose of the article is to determine the most effective model from the BERT family for identifying destructive content in social media. A comparison of the five most well-known BERT models for identifying destructive content was made. For this purpose, a text corpus was created from social media materials (CM), supplemented with Nazi content prohibited for distribution in the Russian Federation from the Federal List of Extremist Materials. The structure of the text data classifier based on the deep artificial neural network BERT is presented and its operation at each stage is described. The search for the most effective method of preprocessing texts was carried out. The efficiency of various heads of classifiers based on the BERT transformer is evaluated. The influence of BERT model retraining is estimated and the effectiveness of its application with the calculation of perplexy is proved. Comparative tables of classifiers' work at each stage of the study are presented. The most effective architecture of the classifier based on the BERT transformer has been found, which performs the task of identifying destructive content with an accuracy of 96.99%.

Keywords: destructive text content, extremism, social media, BERT transformer, classification accuracy.

Submitted 23.08.2022

**Information about the authors**

**Vladimir A. Minaev** – Dr. Sc. (Technical), Professor, Professor of the Department of Special Information Technologies, V. Ya. Kikot Moscow University of the Russian Internal Affairs Ministry, e-mail: m1va@yandex.ru
**Alexander V. Simonov** – Post-graduate student of the Information Security Department, Bauman Moscow State Technical University, e-mail: san.siman@yandex.ru