ВЕСТНИК НАЦИОНАЛЬНОГО ИССЛЕДОВАТЕЛЬСКОГО ЯДЕРНОГО УНИВЕРСИТЕТА “МИФИ”, 2021, том 10, № 6, с. 529–539

ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА

УДК 004.912

СРАВНЕНИЕ ТОЧНОСТЕЙ МЕТОДОВ НА ОСНОВЕ ЯЗЫКОВЫХ И ГРАФОВЫХ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ ДЛЯ ОПРЕДЕЛЕНИЯ ПРИЗНАКОВ АВТОРСКОГО ПРОФИЛЯ ПО ТЕКСТАМ НА РУССКОМ ЯЗЫКЕ

© 2021 г. А. Г. Сбоев[1,2,*], Р. Б. Рыбка[1], И. А. Молошников[1], А. В. Наумов[1], А. А. Селиванов[1]

[1] НИЦ “Курчатовский институт”, Москва, 123182 Россия

[2] Национальный исследовательский ядерный университет “МИФИ”, Москва, 115409, Россия

**e-mail: sag111@mail.ru*

Поступила в редакцию 08.02.2022 г.
После доработки 10.02.2022 г.
Принята к публикации 22.02.2022 г.

В работе анализируются эффективность методов авторского профилирования-определения пола и возрастной группы автора текста, а также наличия в тексте признаков намеренного искажения по полу, возрасту или стилю. При этом используется собранный в ходе исследования с использованием методов краудсорсинга корпус, наиболее представительный в настоящее время на русском языке с разметкой для задачи определения наличия имитации характеристик авторского профиля. Помимо классификационных алгоритмов, основанных на методах опорных векторов и глубоких нейронных сетях, обученных по принципу языковых моделей, мы исследуем применение графовых нейронных сетей, в которых текст представляется набором морфо-синтаксических признаков. В результате проведенного исследования оценен текущий уровень точности решения для задач определения признаков авторского профиля, знание которого в совокупности с собранным набором данных, на котором они получены, может быть полезно в качестве ориентира для апробации новых методов машинного обучения.

Ключевые слова: авторское профилирование, искажение текстов, анализ текстов, машинное обучение, нейронные сети, графовые нейронные сети

DOI: 10.56304/S2304487X21060109

1. ВВЕДЕНИЕ

Интерес к задачам авторского профилирования объясняется безусловным влиянием характеристик авторов (пол, возраст, уровень образования, темперамента и т.д.) на лингвистические особенности авторского текста [1], что позволяет ставить задачу определения по текстам характеристик их авторов. Решение этой задачи актуально, в частности, для построения эффективных инструментов социального мониторинга пользователей на основе интернет-информации, включая тех, для которых не доступны данные их авторского профиля [4]. В ряде случаев авторы текстов могут сознательно искажать признаки письменной речи с целью имитации речи лица противоположного пола и/или другой возрастной группы, что в значительной степени осложняет задачу их определения и требует создания специальных размеченных корпусов.

Сокрытие автором своих настоящих признаков – частный случай лжи в тексте. В работах [1–3] рассматривают выявление лжи или обманчивых утверждений в тексте. Здесь зачастую рассматривают наборы данных (корпуса) размеченные с помощью краудсорсинга: корпус отзывов (Ott Deceptive Opinion Spam corpus [5]), корпус новостей и записей из социальных сетей (LIAR fake news dataset [6]), корпус авторских текстов на заданную тематику (Open Domain Deception Dataset [7]). Развитию размеченных наборов корпусов по авторскому профилю способствовало проведение крупных международных конференций и соревнований, в которых на протяжении последних лет затрагиваются темы намеренного изменения признаков авторского профиля: определение признаков изменения стиля PAN в 2017–2021 [8], сокрытия авторства PAN в 2016–2018 [9], FIRE 2019 [10].

Перечень наиболее популярных методов, достигающих лучшие точности на этих задачах, включает различные модели машинного обучения (такие как: SVM, Naive Bayes, GBM и др.) и их ансамбли, нейронные сети (MLP, LSTM), а также

современные языковые модели на базе архитектуры трансформер (BERT). В качестве используемых полезных признаков отмечаются словари стоп-слов, лексические словари и эмотиконы, частеречные признаки, *n*-граммы слов, а также различные векторные представления слов. В работе [11] показано успешное применение классического классификационного метода на базе SVM и векторизацией слов посимвольным TF-IDF. В работе [12] при решении задачи определения изменения стиля в лучшем подходе в рамках конференции PAN'21 [8] использовали предобученную языковую модель BERT [13] с последующей настройкой на доступных тренировочных данных. В другой работе показали эффективность применения эволюционных методов для обучения сети прямого распространения [14].

Для русского языка в литературе представлено несколько корпусов с текстами, содержащими признаки авторского профиля [1, 15–17], использованные нами в настоящей работе. Методологическая основа для создания корпусов была апробирована в работах [18–20] при создании корпусов RusPersonality и Gender Imitation.

В работе [15] был представлен корпус с текстами, содержащими в себе 1182 текста, которые написаны как без изменения признаков пола, так и с признаками имитации пола и стиля (корпус Gender imitation (GI)). Всего в корпусе содержится по 394 текста по каждому типу имитации. Всего авторов в корпусе: 125 мужчин и 269 женщин. Тексты были написаны по нескольким предварительно заданным тематикам. Вторая часть этого корпуса собиралась с помощью краудсорсинга (GI_cs) и содержала в себе 9612 текстов (из них 3204 с имитацией пола и 3204 с имитацией стиля) от 1161 мужчин и 2043 женщин. В работе [16] этот корпус был расширен с помощью краудсорсинга частью GI_cs_free_theme, содержащей еще 1605 текстов (из них 535 с имитацией пола и 535 с имитацией стиля) от 480 мужчин и 1125 женщин. При этом в задании для краудсорсинга не было указаний на какую тему писать текст. В работе [17] дополнительно рассматривался вопрос определения имитации возраста в тексте. Исследование проводилось на корпусе Age-Imitation-Crowdsource corpus (AI_cs), собранном с помощью краудсорсинга. Корпус включал 12180 текстов от 3302 уникальных авторов, разделенных на 4 возрастные группы: 20–30, 30–40, 40–50 и Other с 5049, 3339, 1680 и 2112 текстами соответственно. Наилучшие результаты по классификации признаков авторского профиля для русскоязычных корпусов текстов демонстрируют подходы на базе технологий машинного обучения: методы на базе машины опорных векторов, градиентного бустинга, а также нейросетевые модели на основе сверточных, рекуррентных и графовых слоев. Представленные исследования ранее были выполнены на отдельных корпусах, собираемых под конкретную задачу [21].

В данной работе для анализа эффективности методов решения задачи авторского профилирования-определения пола, возрастной группы автора текста, а также наличия в тексте признаков намеренного искажения текста по полу, возрасту или стилю составлен корпус текстов русскоязычных авторов, агрегирующий доступные наборы данных, включающих случаи намеренного искажения оригинального авторского стиля, а также характеристик авторского профиля: пола и возраста.

Состав корпуса представлен в разделе 2. Наборы данных в составе корпуса собирались по лингвистическим методикам с применением средств краудсорсинга [15–17]. Предложенный совокупный корпус в настоящее время является наиболее представительной коллекцией текстовых данных для данной задачи. На его основе проведено исследование методов эффективного определения наличия искажения разного вида, а также характеристик авторского профиля в сложных случаях. Сравнивались существующие различные методы решения этой задачи, в т.ч. базовые, основанные на применение частотного кодирования (TF–IDF) и машины опорных векторов (SVM), а также современные методы на базе нейросетевых моделей. В рамках нейросетевого подхода рассмотрен класс методов на основе предобученных языковых моделей, а также на базе графового подхода, когда текст представляется векторами морфологических признаков и матрицей синтаксической связности слов текста. Графовая модель на базе морфо-синтаксических признаков является контекстно-независимой и обладает существенно более простой архитектурой по сравнению с моделями языкового типа, что позволяет проведение большего числа экспериментов по подбору гиперпараметров в процессе обучения для решаемой задачи. Подробное описание графовой модели и используемых методов для сравнения представлено в разделе 3. Общая методология проводимых исследований представлена в разделе 4, а в разделе 5 приводятся результаты сравнительных экспериментов.

2. КОРПУС ТЕКСТОВ С РАЗМЕТКОЙ ПО ПРИЗНАКАМ ПРОФИЛЯ АВТОРОВ

Основу используемого в данной работе набора текстовых данных составляют корпуса текстов: GI_cs, GI_cs_free_theme, AI_cs [15–17]. Данные корпуса собраны с использованием краудсорсинговой платформы Яндекс.Толока, а при их создании использовались лингвистические методы, применяемые для формирования корпусов меньшего объема при очном опросе респондентов (корпус GI).

При создании GI_cs респонденты должны были написать 3 текста:

1) так как они бы писали от себя (метка класса текста: no_gender_imitation);

2) написать текст от лица противоположного пола (with_gender_imitation);

3) написать текст исказив свой обычный стиль написания (with_style_imitation).

Все три текста должны были быть написаны на одну из заданных тем:

– анкета на сайт знакомств с описанием себя и желаемого партнера;

– попытка уговорить незнакомого человека приехать для личной встречи;

– негативный отзыв о туристической поездке.

При создании корпуса GI_cs_free_theme задание было аналогичным, но тема заранее не задавалась.

В процессе создания AI_cs респонденты должны были написать 3 текста:

1) от себя (метка класса текста: no_age_imitation);

2) написать текст от человека сильно моложе себя (young_age_imitation);

3) написать текст от человека сильно старше себя (older_age_imitation)

Все три текста должны были быть написаны на одну из заданных тем:

– попытка убедить произвольного слушателя встретиться с респондентом у него дома;

– рассказ о каком-то запоминающемся событии/приобретении/слухе или чем-то еще, что должно понравиться слушателю;

– рассказ о себе или о ком-то другом, с целью понравиться слушателю и завоевать его расположение.

При выполнении задания респонденты платформы краудсорсинга предварительно указывали пол, возраст, сколько полных лет. При подготовке корпуса сверялись признаки пола и возраста авторов, указанные в профиле автора краудсорсинговой платформы и в задании. Профили с несовпадающими признаками удалялись из выборки.

Дополнительно проводилась автоматическая проверка на заимствования с использованием API сайта text.ru (дата обращения: 19 января 2022) и ручная проверка наиболее длинных и коротких текстов.

В итоговом корпусе от одного автора могло быть более трех документов, т.к. пользователи имели возможность выполнять задание несколько раз и для разных тем.

При этом каждый текст составленного корпуса содержал следующие метки классов:

– gender – бинарная метка пола (“female” – женщины, “male” – мужчины);

– age_group – метка возрастной группы (возрастные рамки указаны включительно: “–19” – возраст до 19 лет, включительно, “20–29” – возраст от 20 до 29 лет, “30–39” – возраст от 30 до 39 лет, “40–49” – возраст от 40 до 49 лет, “50+” – возраст больше или равен 50 годам);

– age_imitation – наличие имитации пола (“no_age_imitation” – от себя, “younger” – написать текст от человека сильно моложе себя, “older” – написать текст от человека сильно старше себя, “None” – неприменимо, указывается для документов из части GenderImitation);

– gender_imitation-имитация текста автора противоположенного пола (“no_gender_imitation” – текст написан от себя, “with_gender_imitation” – текст от лица противоположного пола, “None” – неприменимо, указывается для документов из части AgeImitation);

– style_imitation – имитация стиля (“no_style_imitation” – текст написан от себя, “with_style_imitation” – текст написан с искажением своего обычного стиля написания, “None” – неприменимо, указывается для документов из части AgeImitation);

– no_imitation – отсутствие какой-либо имитации в тексте, возможные значения (“no_im” – текст написан от себя, “with_any_im” – текст написан с каким-либо намеренным искажением пола, возраста или стиля).

Для проведения последующих экспериментов объединенный корпус был сбалансирован по числу авторов женского и мужского пола. Сбалансированный корпус был разделен по авторам равномерно случайным способом на три части: 70% тренировочная, 10% валидационная, 20% тестовая. Таким образом, возможность использования текстов одного автора в разных частях исключается. Тренировочная и валидационная части используются для построения моделей, тестовая для оценки. В табл. 1 представлены сведения о составленном корпусе.

3. МЕТОДЫ

3.1. Графовая сеть с вниманием

Данная архитектура (далее GAM – Graph Attention Model) состоит из комбинации полносвязных слоев и блоков нейронной сети, которые в процессе работы вычисляют коэффициенты значимости объектов x_i входной последовательности X. Особенность данной архитектуры в возможности использования помимо линейной последовательности объектов информации об их связях, заданной в виде матрицы связности M. В данной работе в качестве объектов используются вектор морфологических признаков слов тек-

Таблица 1. Состав собранного корпуса

		Тренировочная часть	Валидационная часть	Тестовая часть
Количество документов		9564	1320	2564
Длина текстов в символах	min	197	200	199
	mean	500.5	520.8	515.6
	max	2984	2809	3981
Количество документов авторов разного пола	жен.	4860	633	1290
	муж.	4704	687	1274
Количество документов авторов по возрастным группам	младше 19	813	117	195
	20–29	4188	570	1131
	30–39	2697	339	683
	40–49	1194	162	351
	старше 50	672	132	204
Количество документов по типу имитации	пола	1215	156	292
	возраста	3946	568	1126
	стиля	1215	156	292
	без имитации	6376	880	1710

ста, а связность определяется синтаксическим деревом (направленным графом). Дополнительно в начало текста ставится слово “cls”, используемое на выходе GAM для определения класса текста. Вначале выполняется линейное преобразование объектов x_i с получением Z – матрицы преобразованных входных объектов. Далее проводится преобразование Z последовательностью из L блоков вычисления коэффициентов значимости (далее блоки внимания). На каждом l блоке внимания ($l \in L$) проводится:

1) расчет векторных представлений слов S с учетом синтаксического окружения M: $S^l = M \cdot Z^l$, где Z^l – вход для l-го блока внимания;

2) получение S'^l – матрицы, характеризующей объекты в S^l с учетом коэффициентов значимости, рассчитываемых в слое Трансформер на основе алгоритма Multi Head Attention из работы [22];

3) формирование входа для $l + 1$ блока внимания. Исследовалось два варианта, либо $Z^{l+1} = S'^l$, либо $Z^{l+1} = BN(S'^l + S^l)$. Во втором случае BN (Batch Normalization) является дополнительным слоем для нормализации по результирующим компонентам вектора объектов в анализируемом наборе текстов.

Таким образом, каждый следующий блок расчетов коэффициентов внимания графовой нейронной сети учитывает больше структурной информации при формировании векторного представления графа. Первый объект результирующей Z^{L+1}, соответствующий закодированному слову “cls”, подается на вход полносвязному слою с функцией активации SoftMax с размерностью, равной числу классов решаемой задачи.

Итоговое число параметров GAM-сети можно оценить по следующей формуле:

$$N_{weights} = (7H_m^2 + 2H_m H_{ff})L + H_m(N_{feat} + N_y), \quad (1)$$

где H_m, H_{ff} – число нейронов в слоях сети в полносвязных слоях в составе Трансформер слоя блока внимания; L – число блоков внимания; N_{feat} – размер признакового пространства векторного представления входных данных; N_y – число классов в задаче классификации.

В разработанной архитектуре сети, помимо параметров H_m, H_{ff}, L, необходимо подобрать:

– число блоков внимания;

– коэффициент dropout [23];

– использовать ли дополнение векторов признаков позиционным кодированием;

– использовать ли сквозные связи (residual) в блоках внимания [24];

– размер батча.

Подбор гиперпараметров сети осуществляется с помощью методов автоматического подбора, которые подробнее рассмотрены в разделе 3.2.

3.2. Алгоритмы подбора гиперпараметров

3.2.1. Tree-Parzen Estimators (TPE). Формирование топологии для решения определенной задачи моделирования на основе данных предполагает выбор конфигурации сети, которая включает в себя совокупность значений гиперпараметров

сети. Для выбора конфигурации был использован адаптированный алгоритм Tree-structured Parzen Estimator (TPE) [25], который позволяет итеративно подобрать конфигурацию сети, оптимальную с точки зрения выбранной метрики в рамках решения задачи.

Алгоритм TPE включает в себя следующие шаги:

1) определение пространства поиска для каждого из свободно настраиваемых параметров сети;

2) определение целевой функции, которая будет оптимизироваться в процессе поиска значений параметров;

3) проведение расчетов для нескольких случайно выбранных комбинаций параметров;

4) сортировка полученных результатов на основе значений целевой функции и разделение по порогу на две группы: та, которая содержит лучшие результаты (×1), и все остальные (×2);

5) расчет плотностей распределения $l(x1)$ и $g(x2)$ на основе ядерной оценки плотности (метода окна Парзена–Розенблатта);

6) случайный выбор множества конфигураций из распределения $l(x1)$, оценка их с точки зрения $l(x1)/g(x2)$, выбор конфигурации, которая соответствует наибольшему ожидаемому улучшению значения целевой функции, расчет целевой функции на основе текущей конфигурации;

7) добавление новых результатов в список результатов из пункта 3;

8) шаги 4–7 повторяются фиксированное число итераций или пока не будет достигнут лимит по времени.

В отличие от алгоритмов [26–28], опирающихся в большей степени на построении архитектуры нейронной сети на основе перебора (и поиска) отдельных блоков топологии и связей между ними. При известной, как у нас топологии сети, поиск на основе разработанного алгоритма с использованием TPE позволяет получить большую точность, чем ручной подбор и за меньшее число итераций, чем алгоритмы, основанные на случайном поиске (Random Search) и переборе (Grid Search).

3.2.2. Population Based Training (PBT). Этот алгоритм основан на эволюционной стратегии [29], которая позволяет контролировать значения гиперпараметров, используемых для обучения моделей. Создается несколько экземпляров модели с разными гиперпараметрами. Каждая модель проходит этап обучения, который может быть ограничен по количеству шагов изменения весов, но их должно быть достаточно для того, чтобы было заметно изменение целевой метрики. После обучения полученные модели оцениваются с использованием валидационного множества. На основе оценок, полученных для каждой модели, принимается одно из двух решений: либо модель считается неперспективной и она удаляется из рассмотрения, либо проводится мутация гиперпараметров, не влияющих на архитектуру модели, и ее обучение продолжается. Экземпляр модели считается неперспективным, если после оценки он находится в нижней квартили. Тогда этот экземпляр заменяется на случайный экземпляр из верхней квартили. Мутация гиперпараметров происходит следующим образом. Для каждого гиперпараметра с вероятностью $P = 0.25$ выбирается новое случайное значение из заранее определенного распределения или списка значений для перебора. С вероятностью $(1 - P)$ старое значение заменяется на ближайшее к нему. Если возможные значения гиперпараметра представлены дискретным списком возможных значений, то выбирается соседнее правое или левое (с равной вероятностью). Если для гиперпараметра задано непрерывное распределение случайных значений, то исходное значение умножается на 0.8 или 1.2 (с равной вероятностью).

Таким образом, конечная модель, получена в результате работы алгоритма, обучена с использованием последовательной комбинации гиперпараметров, выбранных в результате эволюционного отбора.

3.3. Базовая модель

В этой модели для векторизации текста использовался частотный алгоритм TF-IDF по *н*-граммам символов длиной от 3 до 8. Расчет TF-IDF производился на тренировочной части корпуса для каждой задачи (пол, возрастная группа и т.д.) отдельно, без учета примеров с значением целевого класса None (см. раздел 2). В качестве классификатора применялась модель на базе машины опорных векторов с линейным ядром (SVM). Мы использовали численную реализацию базовой модели из библиотеки scikit-learn с гиперпараметрами по умолчанию.

3.4. Метод классификации на основе языковых моделей

В данной работе в качестве языковых мы использовали две модели: RuBERT [30] и RuBERT-tiny [31]. Обе основаны на слоях типа Трансформер. Архитектура RuBERT повторяет архитектуру языковой модели BERT [13]. Настройка весов RuBERT проводилась на основе технологии трансферного обучения, за основу взята предобученная на текстах 104 языков модель Multilingual BERT, которая дообучалась на текстах проекта Wikipedia на русском языке. Сложность модели можно оценить по формуле:

$$param = 4LH_m^2 + 2LH_mH_{ff} + VH_m, \quad (2)$$

где L – число слоев трансформер, H_m, H_{ff} – число нейронов в слоях прямого распространения, V – размер словаря модели.

Количество параметров для RuBERT модели составляет 177M ($L = 12$, $H_m = 768$; $H_{ff} = 3072$; $V =$ = 119 тыс.).

RuBERT-tiny [31] архитектурно соответствует модели BERT со следующими изменениями: размер входного словаря уменьшен с 119 тыс. до 30 тыс. наиболее часто встречающихся слов (токенов) в русском и английском языках, размер слоя векторного представления слов уменьшен с 768 до 312, число слоев трансформер уменьшено с 12 до 3. Модель получена в результате процедуры обучения с использованием выходов уже обученных нейросетевых моделей большего размера, в т.ч.: RuBERT [30], LaBSE [32], Laser [33], и USE [34]. Для этого использовалось значение вектора для служебного токена CLS, добавляемого в начало анализируемого текста. В ходе обучения, помимо восстановления замаскированных токенов, решалась задача минимизации разницы значений вектора токена CLS, получаемого из модели RuBERT-tiny, и значений токенов CLS из моделей RuBERT, LaBSE, LASER и USE. В качестве обучающих данных были использованы следующие корпуса: Yandex Translate (https://translate.yandex.ru/corpus), OPUS-100 [35] и Tatoeba [36]. Количество параметров для RuBERT-tiny модели, рассчитанное по оценке, представленной для RuBERT, составляет 11.5M ($L = 3$, $H_m = 312$; $H_{ff} =$ = 600 ; $V =$ 29 тыс.).

4 ЭКСПЕРИМЕНТЫ

4.1. Предобработка данных

Для проведения экспериментов с собранным датасетом была проведена предобработка данных с помощью библиотеки stanza [37]. Сюда включались следующие процедуры:

– токенизация – выделение из текста предложений, а также деление предложения на отдельные слова и символы пунктуации (токены);

– лемматизация – определение леммы каждого из выделенных слов;

– частеречная разметка – определение части речи для каждого из выделенных слов (поддерживается 17 различных частеречных тэгов, в т.ч. “NOUN”, “VERB” и др.);

– выделение морфологических признаков – разбор выделенных слов на морфологические признаки соответствующей части речи.

Для настройки моделей в составе Stanza использовался корпус SynTagRus[1] [38]. Точность этих моделей на корпусе SynTagRus, рассчитанная ее авторами[2] в соответствии с правилами соревнования CoNLL-2018, составляет: 99.57, 98.86, 97.51, 98.2 и 95.91 для задач: токенизации, выделения предложений, лемматизации, частеречной разметки и выделения морфологических признаков соответственно, по метрике F1.

4.2. Обучение моделей

Исследование выполняется на собранном множестве примеров с фиксированным разделением на тренировочное, тестовое и валидационное подмножества. Методы, описанные в разделе 3, используются для создания моделей решения каждой классификационной задачи отдельно с учетом наличия меток соответствующих классов у исходных объектов – текстов. В качестве задач рассматриваются:

1) задача бинарной классификации текстов по классам пола автора: мужчина или женщина;

2) мультиклассовая классификация текстов по 5 возрастным группам;

3) бинарная задача определения наличия имитации пола: в тексте присутствуют признаки искажения пола автора или нет;

4) бинарная задача определения наличия имитации возраста: в тексте присутствуют признаки искажения возраста автора или нет;

5) бинарная задача определения наличия имитации стиля: в тексте присутствуют признаки искажения стиля автора или нет;

6) бинарная задача определения наличия имитации пола, возраста или стиля.

Для задач 1 и 2 обучение моделей проводится на всем доступном множестве текстов без разделения на случаи отсутствия и наличия имитации. Тестирование выполняется также на общем множестве, но с детализацией на разные случаи. В случае искажения стиля, точно неизвестно кого именно пытался имитировать выполняющий задание пользователь. Возможны случаи, когда выполняя это задание, он писал текст от лица другого пола. Поэтому при создании моделей решения задачи имитации пола, возраста или стиля (задачи № 3, 4 и 5) не используются примеры для других типов имитации. Т.е. при обучении модели определения имитации стиля, не используются примеры с имитацией пола и возраста.

Для GAM-модели используется несколько вариантов гиперпараметров, в том числе:

[1] Корпус SynTagRus: https://universaldependencies.org/treebanks/ru_syntagrus/index.html

[2] Точность моделей stanza: https://stanfordnlp.github.io/stanza/v100performance.html

Таблица 2. Точности определения характеристик авторского профиля

Модель	Определение пола			Определение возрастной группы		
	Все тестовые данные	Тестовые данные с имитацией любого типа	Тестовые данные без какой-либо имитации	Все тестовые данные	Тестовые данные с имитацией любого типа	Тестовые данные без какой-либо имитации
RuBERT	**0.79**	0.73	0.90	**0.40**	0.39	0.42
TinyBERT (Russian)	0.76	0.72	0.84	0.37	0.36	0.37
GAM и PBT	0.75	0.69	0.86	**0.40**	0.40	0.40
TF-IDF + SVM	0.67	0.65	0.73	0.37	0.36	0.37
GAM и TPE	0.74	0.70	0.82	0.34	0.34	0.34
GAM и fixed param	0.70	0.66	0.77	0.32	0.33	0.31
Baseline	0.49	0.49	0.49	0.29	0.28	0.31

1) выбранные в литературе для решения задачи определения пола автора на корпусе без искажения (далее fixed param),

2) подобранные алгоритмы TPE. При этом подбор осуществлялся отдельно для каждой из решаемых задач,

3) настроенные с использованием алгоритма PBT. Как и в случае с TPE, настройка гиперпараметров проводилась для каждой задачи отдельно.

Для сравнения используется набор простейший классификатор, который определяет класс примера с учетом представительности каждого класса в обучающей выборке (далее Baseline).

4.3. Конфигурация подбора гиперпараметров для GAM-модели

Подбор гиперпараметров производился индивидуально для каждой из шести задач. Целевая функция – максимизация классификационной метрики F1 с взвешиванием пропорциональным количеству примеров каждого класса (далее F1-weighted). Каждый алгоритм работал не более четырех часов для решения каждой из задач. При этом использовалась машина со следующей конфигурацией: GPU: 4x Nvidia Tesla V100 с 32 Гб памяти каждая; CPU: Intel Xeon Processor E5-2660 v4 (по 8 процессорных ядер на 1 GPU); RAM: 384 Гб. Одновременно на машине запускалось 4 параллельных экземпляра модели с разными гиперпараметрами, работающими на отдельных GPU. Максимальное число вариантов для перебора 100, максимальное число циклов (итераций для PBT или эпох для TPE) 100, для TPE использовался ранний останов после 10 эпох.

PBT и TPE производили поиск по следующим гиперапараметрам: “batch_size”: равномерно, от 2 до 32 с шагом 2; “learning_rate”: loguniform от 0.0001 до 0.01; “neurons_1_multi”: равномерно, от 4 до 32 с шагом 2; “neurons_2”: равномерно, от 4 до 128, с шагом 4; “n_blocks”: равномерно, от 1 до 8 с шагом 1; “n_heads”: равномерно от 1 до 10 с шагом 1; “dropout”: равномерно, от 0.0, до 0.7 с шагом 0.1; “use_residual”: выбор из [True, False]; “add_positional_encoding”: выбор из [True, False].

При использовании PBT в рамках одного запуска менялись параметры: “learning_rate”: логарифмическое распределение, от 0.0001 до 0.01 шкале; “batch_size”: равномерно, от 2 до 32 с шагом 2; “perturbation_interval” = 1.

Для всех запусков сохранялись лучшие веса и оценка модели на тестовом множестве делалась с лучшими весами.

Работы с подбором и настройкой гиперпараметров осуществлялась на базе библиотеки Ray Tune [39].

5. РЕЗУЛЬТАТЫ

В табл. 2, 3 представлены точности решения задачи определения пола и возраста по метрике F1-weighted.

Лучшие точности определения признака пола автора текста на всей части тестовых данных демонстрирует модель на основе языковой модели RuBERT. Предложенная GAM-модель с настройкой гиперпараметров в процессе обучения на базе алгоритма PBT демонстрирует ухудшение точности на 0.04. В задаче определения возрастной группы автора текста на всей части тестовых данных обе эти модели показывают сравнимые результаты с опережением базовой точности, рассчитанной при выборе класса текста из вероятного распределения меток классов решаемой задачи, на 0.11%. При одинаковой точности GAM-модель обладает меньшей сложностью: 52 тыс. ве-

Таблица 3. Точности определения наличия признаков имитации в текстах

Модель	Имитация пола	Имитация возраста	Имитация стиля	Любая имитация
RuBERT	0.66	**0.82**	**0.77**	**0.80**
TinyBERT (Russian)	0.66	0.78	0.76	0.77
GAM и PBT	**0.68**	0.65	0.67	0.68
TF-IDF + SVM	0.62	0.72	0.71	0.71
GAM и TPE	0.63	0.62	0.66	0.67
GAM и fixed param	0.60	0.62	0.67	0.68
Baseline	0.56	0.32	0.57	0.55

сов GAM-модели, и 11.5M и 177M весов у TinyBERT (Russian) и RuBERT-моделей, соответственно. В обеих задачах применение PBT-алгоритма для GAM-модели позволяет улучшить точность решения задач по сравнению с параметрами подобранными TPE-алгоритмом и параметрами, взятыми из литературы (fixed param), в среднем на 0.03 и 0.06 соответственно.

Текущая постановка задачи определения возраста отличается от предыдущих опубликованных работ, представляющих результаты исследования только на части текстов Age Imitation Crowdsource: 1) не используется балансировка по возрастным группам, 2) количество групп увеличено до 5 с добавлением текстов автором с возрастом меньше 19 лет и старше 50 лет.

В табл. 3 представлены полученные точности по задаче определения наличия имитации характеристик авторского профиля.

GAM-модель с обучением на базе PBT демонстрирует лучшие точности при определении наличия в тексте признаков имитации пола. В остальных случаях определения наличия имитации модели на RuBERT показывают лучшие точности. Это можно объяснить использованием авторами текстов другого набора слова при намеренной имитации стиля\возраста, а также большей сложностью классификационных моделей. В случае искажения пола ключевыми являются морфо-синтаксические характеристики, что отражается в эффективности работы GAM модели. Результирующая GAM модель для данной задачи состоит из 164 тыс. весов.

6. ЗАКЛЮЧЕНИЕ

Проведенные исследования методов авторского профилирования на составленном наиболее крупном русскоязычном корпусе позволили сопоставить точности решения задач авторского профилирования и определения наличия имитации в текстах по полу, возрасту, и стилю с использованием различных по сложности алгоритмов и методов настройки их параметров. Показано, что наиболее эффективными по совокупности задач определения возрастной группы и признаков наличия имитации пола являются алгоритмы на основе языковых моделей, им незначительно (в пределах 4%) уступает нейросетевой алгоритм, основанный на графовых сверточных слоях, учитывающий при анализе структурные признаки текстов, выраженные синтаксическими деревьями предложений. Этот алгоритм обладает существенно меньшей сложностью по сравнению с языковыми моделями, однако для задач определения возрастной группы и признаков наличия имитации пола оказывается сравнимым по точности. При настройке с помощью PBT он показывает лучший результат для задачи имитации пола, что объясняется не значимостью словарных признаков в условиях примерно равной представительности в выборке примеров с имитацией и без, и как следствие возникающей при этом значимостью морфологических и структурных (синтаксических) признаков. Таким образом проделанная работа очерчивает применимость графовых моделей к таким задачам, в которых именно структура текста является основным характерным признаком. Раскрытие потенциала графовых моделей для таких задач является направлением последующих работ.

БЛАГОДАРНОСТИ

Исследование выполненно при поддержке гранта РФФИ № 18-29-10084 (мк). Работа была выполнена с использованием оборудования центра коллективного пользования “Комплекс моделирования и обработки данных исследовательских установок мега-класса” НИЦ “Курчатовский институт”, http://ckp.nrcki.ru/.

СПИСОК ЛИТЕРАТУРЫ

1. *Gröndahl T., Asokan N.* Text analysis in adversarial settings: Does deception leave a stylistic trace? // ACM Computing Surveys (CSUR). 2019. V. 52. №. 3. P. 1–36.
2. *Barsever D., Singh S., Neftci E.* Building a Better Lie Detector with BERT: The Difference Between Truth and Lies //2020 International Joint Conference on Neural Networks (IJCNN). IEEE. 2020. P. 1–7.

3. *Levitan S.I.* Deception in spoken dialogue: Classification and individual differences. Columbia University, 2019.
4. *Smetanin S.* The applications of sentiment analysis for Russian language texts: Current challenges and future perspectives // IEEE Access, 2020. V. 8. P. 110693–110719.
5. *Ott M., Choi Y., Cardie C., Hancock J.T.* Finding deceptive opinion spam by any stretch of the imagination // arXiv preprint arXiv:1107.4557, 2011.
6. *Wang W.Y.* "Liar, Liar pants on Fire": A new benchmark dataset for fake news detection // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), Vancouver, Canada, 2017. P. 422–426.
7. *Pérez-Rosas V., Mihalcea R.* Experiments in open domain deception detection // Proceedings of the 2015 conference on empirical methods in natural language processing. 2015. P. 1120–1125.
8. *Bevendorff J., Chulvi B., Sarracén G., Kestemont M., et al.* Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection // International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, Cham, 2021. P. 419–431.
9. *Potthast M. et al.* Overview of the Author Obfuscation Task at PAN 2018: A New Approach to Measuring Safety //CLEF (Working Notes), 2018.
10. *Rangel F. et al.* Overview of the track on author profiling and deception detection in arabic // Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. In: CEUR-WS. org., Kolkata, India, 2019.
11. *Siagian M. et al.* DBMS-KU Approach for Author Profiling and Deception Detection in Arabic // Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019), 2019.
12. *Zhang Z. et al.* Using Single BERT For Three Tasks Of Style Change Detection // CLEF 2021 Labs and Workshops, Notebook Papers. CEUR-WS.org., 2021.
13. *Devlin J. et al.* Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805, 2018.
14. *Safara F. et al.* An author gender detection method using whale optimization algorithm and artificial neural network // IEEE Access, 2020. V. 8. P. 48428–48437.
15. *Sboev A. et al.* Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception // Procedia computer science, 2018. V. 123. P. 417–423.
16. *Sboev A. et al.* A gender identification of text author in mixture of Russian multi-genre texts with distortions on base of data-driven approach using machine learning models // AIP Conference Proceedings. AIP Publishing LLC. 2019. V. 2116. № 1. P. 270006.
17. *Sboev A. et al.* To the question of data-driven identification of author's age for Russian texts with age deceptions using machine learning // Journal of Physics: Conference Series. IOP Publishing. 2019. V. 1205. № 1. P. 012049.
18. *Litvinova T.A., Zagorovskaya O.V., Litvinova O.A.* Russian text corpora for deception detection studies // International Journal of Open Information Technologies, 2017. V. 5. № 11. P. 58–63.
19. *Litvinova O. et al.* Deception detection in Russian texts // Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017. P. 43–52.
20. *Litvinova T. et al.* "Ruspersonality": A Russian corpus for authorship profiling and deception detection // 2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT). IEEE, 2016. P. 1–7.
21. *Sboev A. et al.* A Neural Network Model to Include Textual Dependency Tree Structure in Gender Classification of Russian Text Author // Advanced Technologies in Robotics and Intelligent Systems. Springer, Cham, 2020. P. 405–412.
22. *Vaswani A. et al.* Attention is all you need // Advances in neural information processing systems, 2017. P. 5998–6008.
23. *Srivastava N., Hinton G., Krizhevsky A. et al.* Dropout: a simple way to prevent neural networks from overfitting // The journal of machine learning research, 2014. V. 15. № 1. P. 1929–1958.
24. *Kipf T.N.* Deep learning with graph-structured representations, 2020.
25. *Bergstra J. et al.* Algorithms for hyper-parameter optimization //Advances in neural information processing systems, 2011. V. 24.
26. *Jin H., Song Q., Hu X.* Auto-keras: An efficient neural architecture search system //Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019. P. 1946–1956.
27. *Weill C. et al.* Adanet: A scalable and flexible framework for automatically learning ensembles // arXiv preprint arXiv:1905.00080, 2019.
28. *Miikkulainen R. et al.* Evolving deep neural networks // Artificial intelligence in the age of neural networks and brain computing. Academic Press, 2019. P. 293–312.
29. *Jaderberg M. et al.* Population based training of neural networks // arXiv preprint arXiv:1711.09846, 2017.
30. *Kuratov Y., Arkhipov M.* Adaptation of deep bidirectional multilingual transformers for russian language // arXiv preprint arXiv:1905.07213, 2019.
31. Маленький и быстрый BERT для русского языка, 2021. URL: https://habr.com/ru/post/562064/. [дата обращения 10.06.2021].
32. *Feng F. et al.* Language-agnostic BERT sentence embedding / /arXiv preprint arXiv:2007.01852, 2020.
33. *Artetxe M., Schwenk H.* Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond // Transactions of the Association for Computational Linguistics, 2019. V. 7. P. 597–610.
34. *Cer D. et al.* Universal sentence encoder // arXiv preprint arXiv:1803.11175. – 2018.
35. *Zhang B. et al.* Improving massively multilingual neural machine translation and zero-shot translation // arXiv preprint arXiv:2004.11867, 2020.
36. *Tiedemann J.* Parallel data, tools and interfaces in OPUS // Lrec, 2012. V. 2012. P. 2214–2218.

37. *Qi P. et al.* Stanza: A Python natural language processing toolkit for many human languages // arXiv preprint arXiv: 2003.07082, 2020.
38. *Droganova K., Lyashevskaya O., Zeman D.* Data conversion and consistency of monolingual corpora: Russian UD treebanks // Proceedings of the 17th international workshop on treebanks and linguistic theories (tlt 2018), 2018. V. 155. P. 53–66.
39. *Liaw R. et al.* Tune: A research platform for distributed model selection and training // arXiv preprint arXiv:1807.05118, 2018.

Vestnik Nacional'nogo Issledovatel'skogo Yadernogo Universiteta "MIFI", 2021, vol. 10, no. 6, pp. 529–539

Comparison of the Accuracies of Methods Based on Language and Graph Neural Network Models for Determining Author Profile Features from Russian Texts

A. G. Sboev[a,b,#], R. B. Rybka[a], I. A. Moloshnikov[a], A. V. Naumov[a], and A. A. Selivanov[a]

[a] *National Research Center Kurchatov Institute, Moscow, 123182 Russia*

[b] *National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, 115409 Russia*

[#] *e-mail: sag111@mail.ru*

Received February 8, 2022; revised February 10, 2022; accepted February 22, 2022

Abstract—The efficiency of methods of author profiling: determining the gender and age group of the author of the text, as well as the presence of deliberate distortion of the features of the text by gender, age or style has been analyzed. The corpus used in this work consists of various crowdsourced datasets. This is the most representative corpus of Russian-language texts containing markup for this task in Russian. In addition to classification algorithms based on support vector machines and deep language models, the use of graph neural networks in which text is represented by a set of morphosyntactic features is explored. The study allows us to estimate the current level of accuracy in solving the problems of determining the features of the author's profile. The resulting accuracy and collected dataset can be useful as a benchmark for testing new machine learning methods.

Keywords: author profiling, deceptive text, text analysis, machine learning, neural networks, graph neural networks

DOI: 10.1134/S2304487X21060109

REFERENCES

1. Gröndahl T., Asokan N. Text analysis in adversarial settings: Does deception leave a stylistic trace? *ACM Computing Surveys (CSUR),* 2019, vol. 52, no. 3, pp. 1–36.
2. Barsever D., Singh S., Neftci E. Building a Better Lie Detector with BERT: The Difference Between Truth and Lies.*2020 International Joint Conference on Neural Networks (IJCNN). IEEE,* 2020, pp. 1–7.
3. Levitan S.I. *Deception in spoken dialogue: Classification and individual differences. – Columbia University,* 2019.
4. Smetanin S. The applications of sentiment analysis for Russian language texts: Current challenges and future perspectives. *IEEE Access,* 2020. vol. 8, pp. 110693–110719.
5. Ott M., Choi Y., Cardie C., Hancock J. T. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*, 2011.
6. Wang W.Y. "Liar, Liar pants on Fire": A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers),* Vancouver, Canada, 2017, pp. 422–426.
7. Pérez-Rosas V., Mihalcea R. Experiments in open domain deception detection. *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 1120–1125.
8. Bevendorff J., Chulvi B., Sarracén G., Kestemont M. et al. Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection. *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, Cham, 2021. pp. 419–431.
9. Potthast M. et al. Overview of the Author Obfuscation Task at PAN 2018: A New Approach to Measuring Safety.*CLEF (Working Notes),* 2018.
10. Rangel F. et al. Overview of the track on author profiling and deception detection in Arabic. *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings.* CEUR-WS. org, Kolkata, India, 2019.

11. Siagian M. et al. DBMS-KU Approach for Author Profiling and Deception Detection in Arabic.*Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019),* 2019.
12. Zhang Z. et al. Using Single BERT For Three Tasks Of Style Change Detection. *CLEF 2021 Labs and Workshops, Notebook Papers.* CEUR-WS.org, 2021.
13. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805,* 2018.
14. Safara F. et al. An author gender detection method using whale optimization algorithm and artificial neural network. *IEEE Access*, 2020, vol. 8, pp. 48428–48437.
15. Sboev A. et al. Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception. *Procedia computer science,* 2018, vol. 123, pp. 417–423.
16. Sboev A. et al. A gender identification of text author in mixture of Russian multi-genre texts with distortions on base of data-driven approach using machine learning models. *AIP Conference Proceedings.* AIP Publishing LLC, 2019, vol. 2116, no. 1, pp. 270006.
17. Sboev A. et al. To the question of data-driven identification of author's age for Russian texts with age deceptions using machine learning. *Journal of Physics: Conference Series*. IOP Publishing, 2019, vol. 1205. no. 1, pp. 012049.
18. Litvinova T.A., Zagorovskaya O.V., Litvinova O.A. Russian text corpora for deception detection studies. *International Journal of Open Information Technologies*, 2017, vol. 5. no. 11, pp. 58–63.
19. Litvinova O. et al. Deception detection in Russian texts. *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 43–52.
20. Litvinova T. et al. "Ruspersonality": A Russian corpus for authorship profiling and deception detection. *2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*. IEEE, 2016, pp. 1–7.
21. Sboev A. et al. A Neural Network Model to Include Textual Dependency Tree Structure in Gender Classification of Russian Text Author. *Advanced Technologies in Robotics and Intelligent Systems.* Springer, Cham, 2020, pp. 405–412.
22. Vaswani A. et al. Attention is all you need.Advances in neural information processing systems, 2017, pp. 5998–6008.
23. Srivastava N., Hinton G., Krizhevsky A. et al. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014, vol. 15, no. 1, pp. 1929–1958.
24. Kipf T.N. *Deep learning with graph-structured representations,* 2020.
25. Bergstra J. et al. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 2011, vol. 24.
26. Jin H., Song Q., Hu X. Auto-keras: An efficient neural architecture search system. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1946–1956.
27. Weill C. et al. Adanet: A scalable and flexible framework for automatically learning ensembles. *arXiv preprint arXiv:1905.00080*, 2019.
28. Miikkulainen R. et al. *Evolving deep neural networks: Artificial intelligence in the age of neural networks and brain computing*, Academic Press, 2019, pp. 293–312.
29. Jaderberg M. et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.
30. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213,* 2019.
31. Small and fast BERT for the Russian language, 2021. URL: https://habr.com/ru/post/562064/. [accessed 10.06. 2021].
32. Feng F. et al. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852,* 2020.
33. Artetxe M., Schwenk H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 2019, vol. 7, pp. 597–610.
34. Cer D. et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175,* 2018.
35. Zhang B. et al. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv: 2004.11867*, 2020.
36. Tiedemann J. Parallel data, tools and interfaces in OPUS. *Lrec*, 2012, vol. 2012. pp. 2214–2218.
37. Qi P. et al. Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*, 2020.
38. Droganova K., Lyashevskaya O., Zeman D. Data conversion and consistency of monolingual corpora: Russian UD treebanks. *Proceedings of the 17th international workshop on treebanks and linguistic theories (tlt 2018),* 2018, vol. 155, pp. 53–66.
39. Liaw R. et al. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118,* 2018.