УДК 004.89

А.А. Соболев, А.М. Федотова, А.В. Куртукова, А.С. Романов, А.А. Шелупанов

Методика определения возраста автора текста на основе метрик удобочитаемости и лексического разнообразия

Описана методика определения возраста автора анонимного текста, написанного на русском языке. Рассмотрены основополагающие работы предметной области и методы классификации: метод опорных векторов, наивный байесовский классификатор, сверточные и рекуррентные нейронные сети, fastText и BERT. Для проведения исследования использовался собственный набор данных, содержащий 1,5 миллиона комментариев пользователей социальных сетей. Отдельные эксперименты посвящены оценке влияния различных методов векторизации текста и фильтрации фотографий пользователей социальных сетей при помощи компьютерного зрения на точность классификации. В результате серии экспериментов, направленных на оценку эффективности использованных методов и отбора информативных признаков, достигнута точность определения возраста автора анонимного текста 83,2%.

Ключевые слова: атрибуция, определение возраста, анализ текста, машинное обучение, нейронные сети, отбор признаков.

DOI: 10.21293/1818-0442-2022-25-2-45-52

В последние годы активно развивается такая область машинного обучения, как обработка естественного языка, ввиду потребности общества в решении прикладных задач, связанных с текстом. Одной из таких практических задач является атрибуция текста – определение пола, возраста, профессии автора текста и авторства в целом [1–3]. В данной статье рассматривается задача идентификации возраста автора текста.

Актуальность и практическая значимость исследования заключается в разработке алгоритмического и программного обеспечения, позволяющего проводить автоматическую фильтрацию контента с целью ограждения несовершеннолетних от запрещенного или шокирующего материала, который имеет возрастной рейтинг «18+». Подобные решения также могут применяться в криминалистике [4, 5]. Проведение авторoведческой экспертизы требует привлечения специалистов в области лингвистики, которые, как правило, проводят исследование вручную. Автоматизированные решения позволят значительно ускорить процесс проведения авторoведческих экспертиз и уменьшить в нем участие специалистов-филологов.

Целью исследования является оценка возможности использования методов, основанных на нейронных сетях (НС) и классических методах машинного обучения, для определения возраста автора анонимного русскоязычного текста. В качестве входных параметров принимаются количественные характеристики образцов текстов. При этом для части образцов авторство известно, и текстам сопоставлена метка возраста автора. Необходимо классифицировать анонимные тексты по возрастным категориям. Таким образом, множеством классов будут являться возрастные категории авторов, а тексты, возраст авторов которых известен, – конечной обучающей выборкой. Решение задачи сводится к получению классификатора, который максимально точно определяет возрастные категории.

Научная новизна исследования заключается в применении ранее не использовавшихся в решении данной задачи для русскоязычных текстов методов: fastText, BERT, а также проведении серии экспериментов, направленных на исследование различных методов векторизации текстов и фильтрации пользовательских фотографий.

Анализ предметной области

Проблеме определения возраста автора текста посвящено сравнительно небольшое количество работ [6–10].

Статья [6] направлена на определение возраста автора при помощи глубоких НС. Ключевая особенность исследования состоит в применении методов компьютерного зрения с целью фильтрации недостоверных данных пользователей о возрасте в социальных сетях. Данный метод позволил сформировать обучающий набор, содержащий публичные сообщения пользователей социальной сети «ВКонтакте» в паре с доподлинными возвратами их авторов. Классификация текстов выполнялась при помощи различных архитектур НС, однако наилучшая точность была достигнута моделью fastText и составила 82%.

В работе [7] использовались 15060 текстов 1260 авторов, публиковавших посты в LiveJournal. Для экспериментов были заданы возрастные группы: 20–30, 30–40 и 40–50 лет. Тексты были распределены по заданным группам в равных пропорциях. В качестве лингвистических признаков текста применялись униграммы и биграммы слов, а также распределение слов по частям речи. Классификация осуществлялась методом опорных векторов (SVM), для оценки использовалась кроссвалидация по 10 фолдам. Точность классификации составила 59,8%.

Авторы статьи [8] выполняли определение возраста и пола автора SMS-сообщений. Всего было использовано 38588 предварительно обработанных текстовых сообщений, написанных носителями английского языка и сингапурскими учащимися,

изучающими английский. В качестве классификаторов авторы применяли наивный байесовский классификатор, SVM и алгоритм J48. Последний позволил получить лучший результат для задачи определения возраста автора – 70,79%.

Авторы работы [9] выполняли определение возраста авторов коротких текстов (средняя длина – 100 слов). Подход основан на признаках читабельности текста. Обучающие наборы включали в себя две возрастные группы: дети до 16 лет и взрослые от 20 лет. Модель, созданная с помощью SVM с AdaBoost, позволила достигнуть *f*-меры, равной 94%.

Задача определения возраста автора текста рассматривалась в рамках соревнования PAN-2016. Участники работали с набором данных, состоящим из 686 текстов на английском и испанском языках. В рамках исследования рассматривались возрастные группы 18–24, 25–34, 35–49, 50–64, 65 лет и старше. Команда, занявшая первое место по итогам соревнования, в рамках предобработки текстов удаляла пустые строки, дубликаты текстов и лишние пробелы. Затем тексты преобразовывались в векторы в виде TF-IDF. Далее тексты классифицировали с помощью SVM. Точность метода для английского языка составила 51%, для испанского – 54% [10].

Авторы статьи [11] в качестве признаков использовали текст и его стиль, а также синтаксические и лексические особенности. Классификация происходила с использованием байесовской полиномиальной регрессии. Тексты разделялись на три возрастные категории: 13–17, 23–27 и 33–47 лет. Максимальная точность 77% достигнута при совместном использовании и синтаксических, и лексических признаков.

В работе [12] исследовалось определение демографических атрибутов пользователей социальной сети Twitter. Авторами был собран набор данных на английском, испанском, французском, немецком и итальянском языках. В качестве признаков для SVM были выбраны триграммы слов. Максимальная точность получена для английского языка – 84%.

Следует отметить, что атрибуция в социальных сетях особенно затруднена, так как текст, как правило, не структурирован, имеет небольшой объем, содержит сленг и шумы. Обучение модели будет недостаточно эффективным, если данные ненадежны, зашумлены или не сбалансированы.

Проблемой также является неграмотность пользователей, которая может быть как преднамеренной, так и непреднамеренной. Основная часть описанных недостатков может быть устранена на этапе предобработки текста.

Исследований, направленных на определение возраста автора русскоязычного текста количественными методами, насколько известно авторам статьи, нет. О недостаточной изученности данной темы также свидетельствует отсутствие размеченных общедоступных наборов данных для русского языка. Поэтому одним из этапов исследования был сбор данных, содержащих тексты и метку возраста автора.

Набор данных

В настоящее время большой популярностью пользуются социальные сети. Важными аргументами в пользу сбора данных из социальных сетей являются широкий доступ для людей различных возрастов и многообразие тем для обсуждения.

В связи с этим можно легко найти сообщения подростков, которые будут обсуждать школу, фильмы, музыку или множество других тем, и обсуждения взрослых людей о работе, быте, а также темы, которые будут пересекаться с теми, которые больше интересны пользователям младшего поколения. Для формирования набора данных выбрана социальная сеть «ВКонтакте», так как она является самой популярной для русскоязычного сообщества и охватывает все возрастные категории граждан, а также содержит множество сообщений различных тематик. В 2021 г. каждый месяц пользователями оставлялось по 399 млн сообщений 23 млн авторов, однако всего у 45,2% авторов указан пол.

Всего для исследований было собрано 1,5 млн комментариев 393 тыс. авторов. В рамках исследования тексты были разделены на две возрастные категории: детскую – тексты, авторам которых до 18 лет включительно, и взрослую – тексты, авторам которых от 21 года до 55 лет.

Следует отметить, что тексты, написанные авторами в возрасте от 18 до 21 года, при обучении моделей не учитывались, так как разделяющая способность моделей на этом интервале неудовлетворительна. Кроме того, это оказывает негативное влияние на конечный результат. Ситуация аналогична интервалу от 27 до 30 лет.

Методика определения автора анонимного текста

Методика определения возраста автора анонимного текста представлена на рис. 1. Рассмотрим этапы более подробно.

Предварительная обработка текстов. Для обучения моделей необходимо преобразовать тексты в векторный вид. Прежде всего, привести все символы в нижний регистр, очистить данные от некоторого «шума», такого как стоп-слова и знаки препинания, произвести замену эмотиконов специальным тегом, поскольку удалять их не имеет смысла – эмотиконы являются неотъемлемой частью общения в сети Интернет. Одним из этапов предварительной обработки данных являются лемматизация и стемминг, которые приводят слова к нормальной форме.

Фильтрация данных с использованием компьютерного зрения. Процесс фильтрации заключался в отсеивании пользователей, у которых возраст на фотографиях не совпадал с возрастом, указанным в профиле.

При фильтрации текстов для каждого пользователя было собрано до 10 фотографий. Если возраст больше чем на половине фотографий совпадает с возрастом, указанным в профиле (±2 года), то возраст пользователя считается достоверным.

Существует множество решений для определения возраста по фото. В данной работе использовалась архитектура VGG-Face, предложенная в исследовании [13]. Данная операция очень затратна, так как НС имеет более 145 млн параметров, поэтому обработка одной фотографии занимает в среднем от 5 до 10 с.

В табл. 1 представлены результаты перекрестной проверки с количеством блоков *k*, равным 5, для данных до и после фильтрации. В качестве метода для преобразования текстов в векторный вид был использован Tokenizer. При использовании данного метода каждое целое число является индексом слова в словаре.

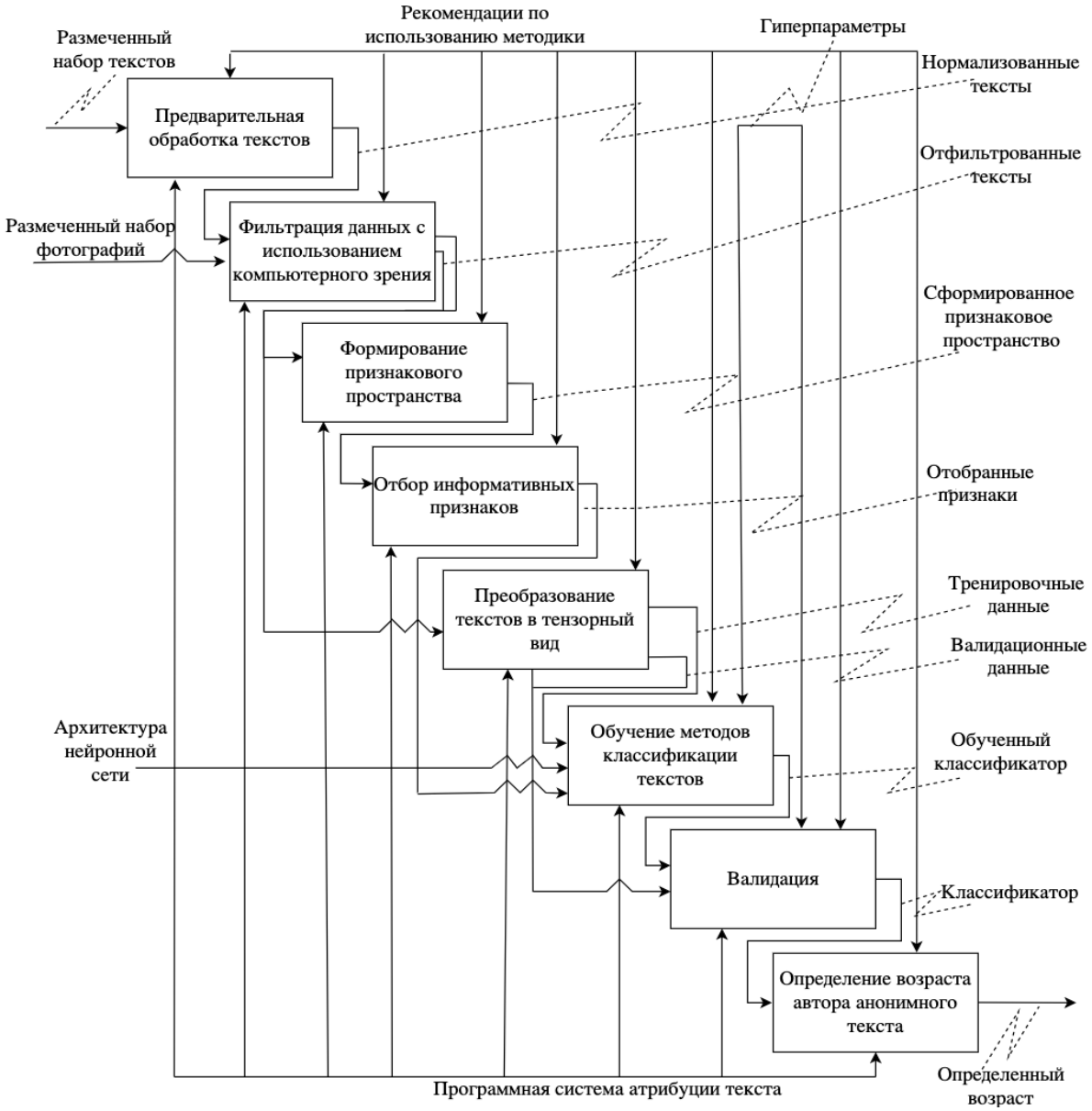

Рис. 1. Методика определения возраста автора анонимного текста

Таблица 1

Результаты с использованием прошедших и не прошедших фильтрацию данных

Метод классификации	До фильтрации		После фильтрации	
	Точность, %	*F*-мера, %	Точность, %	*F*-мера, %
SVM	51,7	66,7	56,2	69,4
Naive Bayes	50,1	66,4	52,7	67,5
Text CNN	65,1	65,1	76,9	76,9
Text RNN	65,2	64,8	74,8	74,3
Text RCNN	64,7	64,6	74,9	74,9

Фильтрация положительно влияет на точность работы моделей, причем точность повышается больше чем на 10%. Лучшая модель по результатам перекрестной проверки – Text CNN, итоговая точность которой составила 76,9%. Из таблицы также видно, что модели с рекуррентными слоями хорошо справляются с поставленной задачей.

Формирование признакового пространства. Было выделено три группы признаков текстов: базовые статистики, метрики удобочитаемости, метрики лексического разнообразия. Базовые статистики содержат следующие статистические данные: количество предложений, слов, уникальных слов, сложных слов, односложных и многосложных слов и др. В

табл. 2 представлены данные признаки для двух возрастных категорий – детской (до 18 лет) и взрослой (18 лет и старше), количественные значения соответствуют характеристикам одного пользовательского сообщения.

Таблица 2

Базовые статистики

Признак	Аудитория	
	детская	взрослая
Кол-во слов	10,9	13,7
Кол-во уникальных слов	9,1	12,6
Кол-во длинных слов	2,8	5,1
Кол-во сложных слов	0,7	1,7
Кол-во простых слов	7,2	10,4
Кол-во односложных слов	2,9	3,9
Кол-во многосложных слов	4,9	8,1
Кол-во символов	59,2	82,1
Кол-во букв	36,9	63,2
Кол-во пробелов	7,8	12,1
Кол-во слогов	16,1	26,8
Кол-во знаков препинания	3,3	3,6

На основе полученных значений отмечено, что женщины оставляют больше сообщений (55,1%), чем пользователи мужского пола (44,8%), а взрослая аудитория пишет более длинные и сложные предложения. Это видно по количеству символов и слов в предложении, употреблению сложных слов в тексте. Знаки препинания и односложные слова одинаково используются и взрослыми, и детьми.

Следующая рассмотренная группа признаков – индексы удобочитаемости. Данные метрики показывают, насколько сложен текст для восприятия.

Тест Флеша–Кинкайда (FKGL) и индекс удобочитаемости Флеша (FRE) используют количество слов, предложений и слогов. Данные индексы изначально были разработаны для английского языка, однако позже были адаптированы для русского путем изменения коэффициентов

$$\mathrm{FKGL} = 0,49\left(\frac{N_{\mathrm{WORDS}}}{N_{\mathrm{SENTENCES}}}\right) + 7,3\left(\frac{N_{\mathrm{SYLLABLES}}}{N_{\mathrm{WORDS}}}\right) - 16,59, \quad (1)$$

$$\mathrm{FRE} = 206,835 - 1,3\left(\frac{N_{\mathrm{WORDS}}}{N_{\mathrm{SENTENCES}}}\right) - 60,1\left(\frac{N_{\mathrm{SYLLABLES}}}{N_{\mathrm{WORDS}}}\right), \quad (2)$$

где N_{WORDS} – количество слов в тексте, $N_{\mathrm{SENTENCES}}$ – количество предложений в тексте, $N_{\mathrm{SYLLABLES}}$ – количество символов в тексте.

Индекс SMOG показывает количество лет обучения, необходимых для понимания текста:

$$\mathrm{SMOG} = 1,1\sqrt{\frac{64,6N_{\mathrm{COMPLEX\_WORDS}}}{N_{\mathrm{SENTENCES}}}} + 0,05, \quad (3)$$

где $N_{\mathrm{COMPLEX\_WORDS}}$ – количество сложных слов.

Индекс удобочитаемости LIX показывает, насколько текст сложен для чтения:

$$\mathrm{LIX} = \frac{N_{\mathrm{WORDS}}}{N_{\mathrm{SENTENCES}}} + \frac{N_{\mathrm{LONG\_WORDS}}}{N_{\mathrm{WORDS}}}, \quad (4)$$

где $N_{\mathrm{LONG\_WORDS}}$ – количество слов длиной более шести букв.

В табл. 3 приведены средние значения индексов удобочитаемости для детской и взрослой аудитории пользователей социальной сети «ВКонтакте»

Таблица 3

Индексы удобочитаемости

Индекс	Аудитория	
	детская	взрослая
Тест Флеша–Кинкайда	0,3	4,27
Индекс удобочитаемости Флеша	97,1	71,7
Индекс SMOG	4,9	8,4
Индекс удобочитаемости LIX	37,4	50,5

Коэффициенты лексического разнообразия показывают богатство словарного запаса автора текста. Самым простым коэффициентом лексического разнообразия является Type-Token Ratio (TTR):

$$\mathrm{TTR} = \frac{N_{\mathrm{LEXEMES}}}{N_{\mathrm{WORDS}}}, \quad (5)$$

где N_{LEXEMES} – количество уникальных слов в тексте.

Коэффициенты Root Type-Token Ratio (RTTR) и Corrected Type-Token Ratio (CTTR) являются модификациями TTR:

$$\mathrm{RTTR} = \frac{N_{\mathrm{LEXEMES}}}{\sqrt{N_{\mathrm{WORDS}}}}, \quad (6)$$

$$\mathrm{CTTR} = \frac{N_{\mathrm{LEXEMES}}}{\sqrt{2N_{\mathrm{WORDS}}}}. \quad (7)$$

В табл. 4 представлены коэффициенты лексического разнообразия для двух категорий пользователей социальной сети «ВКонтакте».

Таблица 4

Коэффициенты лексического разнообразия

Коэффициент	Аудитория	
	детская	взрослая
TTR	0,91	0,96
RTTR	2,75	3,21
CTTR	1,94	2,27

При использовании индекса удобочитаемости Флеша более высокие показатели свидетельствуют о легкости понимания текста. Показатель текста от 100 до 90 баллов говорит о том, что его написал человек, учащийся в 5-м классе. Показатель от 70 до 80 – ученик 7-го класса. Стоит отметить то, что журнал «Time» набирает 62 балла, поэтому результат 71,7 для взрослой аудитории можно считать адекватным.

Также значительно отличается индекс SMOG: для понимания текста детской аудитории необходимо иметь почти 5 классов образования, а взрослой – 8, что сопоставимо с показателями, полученными с помощью индекса удобочитаемости Флеша.

Показатель индекса удобочитаемости LIX для детской аудитории можно интерпретировать как простые тексты, художественную литературу, газетные статьи, а для взрослой аудитории – как тексты средней сложности, журнальные статьи.

Отбор информативных признаков. Для отбора информативных признаков из генеральной сово-

купности признакового пространства был применен дисперсионный анализ. Необходимо получить оценку дисперсии между группами

$$MS_B = \frac{\sum(\bar{X}_j - \bar{X}_G)^2}{k-1} n, \quad (8)$$

где $\overline{X_j}$ – среднее значение группы, $\overline{X_G}$ – общее среднее, k – число групп; n – размер группы.

После оценку дисперсии внутри группы

$$MS_w = \frac{s_1^2 + s_2^2 + s_3^2 + ... + s_n^2}{k}, \quad (9)$$

где s_n – среднеквадратическое отклонение внутри группы. Затем можно вычислить значимость информативных признаков

$$F = \frac{MS_B}{MS_w}. \quad (10)$$

В табл. 5 представлены 10 самых значимых информативных признаков текста.

Таблица 5

Информативные признаки

Признак	*F*
Индекс удобочитаемости Флеша	264,37
Индекс удобочитаемости LIX	193,66
Тест Флеша–Кинкайда	168,49
Индекс SMOG	117,46
Длинные слова	107,01
Буквы	103,92
Слоги	95,74
Сложные слова	93,58
Многосложные слова	91,61
TTR	91,24

Данные признаки будут использоваться при обучении моделей классификации текстов, которые описаны ниже.

Обучение методов классификации текстов. Исходя из проведённого анализа предметной области, можно выделить как несколько зарекомендовавших себя для английского и других языков методов (SVM, наивный байесовский классификатор, сверточные (CNN) и рекуррентные НС (RNN)), так и появившиеся недавно, но демонстрирующие впечатляющие результаты в области анализа теста, вопросно-ответных систем, автоматического перевода, выявления спама и т.д. (fastText, BERT) [14].

Основной задачей SVM является нахождение оптимальной гиперплоскости, способной разделять данные на два класса. Этот процесс происходит за счет максимизации зазора между опорными векторами, являющимися ближайшими точками разных классов в пространстве. Эффективность алгоритма достигается за счет ядрового преобразования, которое отвечает за отображение данных в пространство, в котором разделяющая классы поверхность будет линейной.

Наивный байесовский классификатор основан на теореме Байеса. Данный алгоритм «наивный», так как предполагает независимость признаков. Даже если признаки связаны друг с другом, они вносят независимый вклад в итоговую вероятность. Для определения наиболее вероятного класса в данном алгоритме используется оценка апостериорного максимума, т.е. нахождение вероятностей для множества представленных классов и выбор класса, в котором вероятность максимальна.

Сверточные нейронные сети уже долгое время успешно применяются в области компьютерного зрения. Но в последнее время их также применяют для задач анализа естественного языка, в частности, для классификации текста. Основной принцип работы CNN заключается в работе фильтров, которые распознают определенные особенности данных. Перемещаясь по тексту, фильтр определяет, есть ли искомая характеристика в конкретной части текста. Для получения результата совершается операция свертки, которая является суммой произведений элементов фильтра и матрицы входных сигналов.

Рекуррентные нейронные сети (RNN) – тип нейронных сетей, продемонстрировавший хорошие результаты в обработке последовательных данных, таких как временные ряды и текстовая информация. Данный класс нейронных сетей позволяет нейронам сохранять связи между собой, т.е. передавать информацию от одного шага к другому. Получая информацию на определенном шаге, сеть способна анализировать предшествующую информацию.

В fastText поданные на вход классификатора тексты трансформируются в эмбендинги слов. Далее эмбендинги усредняются, в итоге получается один эмбендинг, который применим ко всему тексту. Результирующий вектор пропускается через классификатор с функцией активации Softmax для расчета итоговых вероятностей.

BERT представляет собой глубокую НС, основанную на кодировщиках Transformer. Каждый уровень кодировщика включает двустороннее внимание. Благодаря этому BERT учитывает контекст с обеих сторон токена, а значит, более точно определяет его смысловое значение.

Для достоверной оценки работы модели необходимо выполнять перекрестную проверку. Также данная проверка позволяет обнаружить переобучение моделей НС.

Валидация. В данной работе использовалась кроссвалидация по *k* блокам. Результат, полученный после перекрестной проверки, можно считать более достоверным. Подбор гиперпараметров является важной частью при обучении моделей, так как от их выбора напрямую зависит способность модели предсказывать правильные ответы.

Ниже приведены гиперпараметры, используемые для обучения CNN, RNN, RCNN, для выбора гиперпараметров использовался поиск по решетке:

– размер мини-выборки – 32;
– скорость обучения – 0,01;
– функция потерь – Binary crossentropy;
– оптимизатор – Adam;

– метрика качества – точность, *F*-мера;

– количество эпох – 100.

Для текстов был выбран размер словаря – 5000 слов, максимальная длина текста – 150 слов. При обучении сверточных НС были взяты ядра – 3, 4, 5.

При обучении fastText установлены: длина входного предложения – 150, эпох – 100, количество негативных слов (негативное сэмплирование) – 5. В случае BERT использовалась мультиязычная базовая модель со стандартными параметрами.

Обучение, валидация и тестирование моделей проводились на одних и тех же наборах данных. Валидационная выборка – 20% от общего числа текстов. Тестовая выборка – 20% от общего числа.

Преобразование текстов в тензорный вид. Помимо Tokenizer, были использованы и другие методы векторизации: быстрое кодирование (One-Hot Encoding), Bag of Words и TF-IDF. В табл. 6 представлены результаты, полученные с использованием перекрестной проверки.

Таблица 6

Результаты с использованием различных методов векторизации

Метод векторизации	Метрика	Text CNN, %	Text RCNN, %
Tokenizer	Точность	76,9	74,9
	F-мера	76,9	74,9
One-Hot Encoding	Точность	54,9	52,9
	F-мера	62,6	60,9
Bag of Words	Точность	55,8	53,3
	F-мера	63,2	61,3
TF-IDF	Точность	61,7	62,7
	F-мера	61,7	62,7

Использовавшийся ранее метод Tokenizer является лучшим при использовании таких НС, как Text CNN и Text RCNN.

Данный подход позволяет выполнить векторизацию текстовых данных, превращая каждый элемент текста либо в последовательность целых чисел (где каждое целое число является индексом токена (лексемы) в словаре), либо в вектор, в котором значение каждого токена может быть бинарным, либо представлено на основании метода Bag-of-Words или TF-IDF. Другие подходы теряют информацию о взаимном расположении слов внутри текста.

Определение возраста автора анонимного текста. Архитектуры НС были изменены путем добавления еще одного входа для информативных признаков текста после конкатенации со слоями обработки текстов. В табл. 7 представлены результаты перекрестной проверки для моделей НС при добавлении информативных признаков, представленных в табл. 4, и без них.

Добавление в модель еще одного входа для обработки информативных признаков улучшает способность модели классифицировать тексты.

В результате проведения серии экспериментов выбраны две лучшие модели – Text CNN и Text RNN с добавлением информативных признаков. Данные модели сравнивались с предобученной мультиязычной моделью BERT, а также с fastText, которая в последнее время показывает хорошие результаты в области обработки естественно-языковых текстов. В табл. 8 представлены результаты предсказаний моделей на тестовой выборке. Тестовая выборка не использовалась при обучении моделей.

Таблица 7

Результаты перекрестной проверки для моделей НС при добавлении информативных признаков

Наличие информативных признаков	Метрика	Text CNN	Text RNN	Text RCNN
Да	Точность, %	76,9	74,8	74,9
	F-мера, %	76,9	74,3	74,9
Нет	Точность, %	78,2	77,3	76,9
	F-мера, %	78,2	77,5	76,8

Таблица 8

Результаты предсказаний моделей на тестовой выборке

Модель	Точность, %	*F*-мера, %
Text RNN + информативные признаки	77,3	77,5
Text CNN + информативные признаки	78,2	78,2
fastText	80,9	80,1
BERT	83,2	83,2

Модели fastText и BERT способны лучше справляться с определением возраста автора анонимного текста. В первую очередь это связано с особым способом приведения текстов в векторное пространство. Модель fastText использует метод Skip-Gram, и негативное сэмплирование. BERT использует позиционное кодирование. Также модель BERT является предобученной, что подразумевает факт того, что она уже обучена на огромном количестве текстов, что значительно повышает итоговую точность.

Заключение

В рамках исследования проведена оценка различных методов определения возраста автора русскоязычного текста, реализована фильтрация пользовательских фотографий с использованием модели VGG-Face. Большое внимание уделялось экспериментальным данным. Русскоязычных корпусов для решения этой задачи нет, поэтому был собран собственный набор данных из социальных сетей. Были проведены эксперименты с добавлением информативных признаков текста в модель. Такой подход увеличивает итоговую точность моделей.

Эксперименты показали, что использование зашумлённых данных без фильтрации снижает точность классификации, так как реальный возраст не совпадает с возрастом, указанным в профилях. Поэтому нельзя доверять результатам методов, оцениваемых на необработанных данных из социальных сетей. Определение возраста по фотографии с помощью VGG-Face и сравнение его с возрастом, указанным в профиле, гарантирует, что пользователь имеет правильно указанный возраст. В этом случае сообщения пользователей можно использовать для обучения модели. Эксперименты показали, что при-

менение процедуры фильтрации позволяет добиться увеличения точности до 12%. Еще один вывод, сделанный из экспериментов с проверенными данными по фотографиям, состоит в том, что пользователи намеренно занижают свой возраст в социальных сетях. Такие действия могут осуществляться в противоправных целях, в частности, для беспрепятственного общения с пользователями, не достигшими 18-летнего возраста. Эксперименты с использованием различных методов векторизации позволяют сделать вывод об особой эффективности выбора подходящего к задаче метода. При классификации с использованием RCNN и метода Tokenizer удалось получить точность на 21% выше, чем с One-Hot Encoding.

Наилучший результат для русскоязычного текста был получен при использовании модели BERT. Точность перекрестной проверки составила 83,2%. Полученная точность метода сравнима с подходами, применяемыми для английского, китайского и других языков, даже с учетом сложности русского языка.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ в рамках базовой части государственного задания ТУСУРа на 2020–2022 гг. (проект № FEWM-2020-0037).

Литература

1. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks / A. Romanov, A. Kurtukova, A. Shelupanov, A. Fedotova, V. Goncharov // Future Internet. – 2021. – No. 1. – URL: https://www.mdpi.com/1999-5903/13/1/3/htm, свободный (дата обращения: 25.03.2021).

2. Kurtukova A. Source Code Authorship Identification Using Deep Neural Networks / A. Kurtukova, A. Romanov, A. Shelupanov // Symmetry. – 2020. – No. 12. – URL: https://www.mdpi.com/2073-8994/12/12/2044/htm, свободный (дата обращения: 04.04.2022).

3. Authorship Attribution of Social Media and Literary Russian-Language Texts Using Machine Learning Methods and Feature Selection / A. Fedotova, A. Kurtukova, A. Romanov, A. Shelupanov // Future Internet. – 2022. – No. 14. – URL: https://www.mdpi.com/1999-5903/14/1/4/htm, свободный (дата обращения: 11.04.2022).

4. Smetanin S. The applications of sentiment analysis for Russian language texts: Current challenges and future perspectives // IEEE Access. – 2020. – Vol. 8. – P. 110693–110719.

5. Sboev A. Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception / A. Sboev, I. Moloshnikov, D. Gudovskikh, A. Selivanov, R. Rybka, T. Litvinova // Procedia computer science. – 2018. – Vol. 123. – P. 417–423.

6. Romanov A.S. Determining the Age of the Author of the Text Based on Deep Neural Network Models / A.S. Romanov, A.V. Kurtukova, A.A. Sobolev, A.A. Shelupanov, A.M. Fedotova // Information. – 2020. – No. 12. – URL: https://www.mdpi.com/2078-2489/11/12/589/htm, свободный (дата обращения: 11.04.2022).

7. Litvinova T. Profiling the age of Russian bloggers / T. Litvinova, A. Sboev, P. Panichev // Conference on Artificial Intelligence and Natural Language. – Springer, Cham, 2018. – P. 167–177.

8. Khdr A.J. Age and Gender Identification by SMS Text Messages / A.J. Khdr, C. Varol // 2018 International Conference on Artificial Intelligence and Data Processing (IDAP). – 2018. – P. 1–5.

9. Pentel A. Automatic Age Detection Using Text Readability Features // EDM. – 2015. – No. 1146. – P. 40–45.

10. Madhulika A. Age and gender identification using stacking for classification / A. Madhulika, T. Gonçalves // CLEF. – 2016. – P. 785–790. – URL: https://dspace.uevora.pt/rdpc/handle/10174/20668, свободный (дата обращения: 12.04.2022).

11. Argamon S. Automatically Profiling the Author of an Anonymous Text / S. Argamon, M. Koppel, J. Pennebaker, J. Schler // Commun. ACM. – 2009. – No. 52. – P. 119–123.

12. Коршунов А. Определение демографических атрибутов пользователей микроблогов / А. Коршунов, И. Белобородов, А. Гомзин // Труды Института системного программирования РАН. – 2013. – № 25. – С. 179–194.

13. Masi I. Deep face recognition: A survey / I. Masi, Y. Wu, T. Hassner, P. Natarajan // 2018 SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). – 2018. – P. 471–478.

14. Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding / J.Devlin, M.W. Chang, K. Lee, K. Toutanova // arXiv preprint arXiv:1810.04805. – 2018. – URL: https://arxiv.org/pdf/1810.04805.pdf, свободный (дата обращения: 13.04.2022).

Соболев Артем Александрович
Студент каф. безопасности информационных систем (БИС) Томского государственного университета систем управления и радиоэлектроники (ТУСУР)
Ленина пр-т, 40, г. Томск, Россия, 634050
Тел.: +7-952-183-06-83
Эл. почта: bingjo-ya@yandex.ru

Федотова Анастасия Михайловна
Студентка каф. БИС ТУСУРа
Ленина пр-т, 40, г. Томск, Россия, 634050
Тел.: +7-923-444-41-25
Эл. почта: afedotowaa@yandex.ru

Куртукова Анна Владимировна
Аспирант каф. комплексной информационной безопасности электронно-вычислительных систем (КИБЭВС) ТУСУРа
Ленина пр-т, 40, г. Томск, Россия, 634050
Тел.: +7-905-991-67-13
Эл. почта: av.kurtukova@gmail.com

Романов Александр Сергеевич
Канд. техн. наук, доцент каф. КИБЭВС ТУСУРа
Ленина пр-т, 40, г. Томск, Россия, 634050
Тел.: + 7 (382-2) 41-34-26
Эл. почта: alexx.romanov@gmail.com

Шелупанов Александр Александрович
Д-р техн. наук, проф., зав. каф. КИБЭВС ТУСУРа
Ленина пр-т, д. 40, г. Томск, Россия, 634050
Тел.: +7 (382-2) 90-71-55, внут. 10-20
Эл. почта: saa@fb.tusur.ru

Sobolev A.A., Fedotova A.M., Kurtukova A.V.,
Romanov A.S., Shelupanov A.A.

Methodology to determine the age of the text's author based on readability and lexical diversity metrics

The article describes the approaches to determining the age of the author of an anonymous text written in Russian. The fundamental works of the subject area are considered, both proven approaches (support vector machine, naive Bayes classifier, convolutional and recurrent neural networks) and modern methods (fastText, BERT) are implemented. The study used its own data set containing 1,5 million comments from social media users. A separate experiment is devoted to assessing the impact on the classification accuracy of various text vectorization methods. As a result of a series of experiments aimed at evaluating the efficiency of the methods used and selecting informative features, a model was obtained that can predict the age of the author of an anonymous text with an accuracy of 83.2%.

Keywords: attribution, age determination, text analysis, machine learning, neural networks, feature selection.
DOI: 10.21293/1818-0442-2022-25-2-45-52

References

1. Romanov A., Kurtukova A., Shelupanov A., Fedotova A., Goncharov V. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. *MDPI Future Internet*, 2021, no. 1. Available at: https://www.mdpi.com/1999-5903/13/1/3/htm (accessed: 25.03.2021).

2. Kurtukova A., Romanov A., Shelupanov A. Source Code Authorship Identification Using Deep Neural Networks. *MDPI Symmetry*, 2020, no. 12. Available at: https://www.mdpi.com/2073-8994/12/12/2044/htm (accessed: 04.04.2022).

3. Fedotova A., Kurtukova A., Romanov A., Shelupanov A. Authorship Attribution of Social Media and Literary Russian-Language Texts Using Machine Learning Methods and Feature Selection. *MDPI Future Internet.* 2022, no. 14. Available at: https://www.mdpi.com/1999-5903/14/1/4 (accessed: 11.04.2022).

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. Sboev A., Moloshnikov I., Gudovskikh D., Selivanov A., Rybka R., Litvinova T. 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.

6. Romanov A.S., Kurtukova A.V., Sobolev A.A., Shelupanov A.A., Fedotova A.M. Determining the Age of the Author of the Text Based on Deep Neural Network Models. *MDPI Information*. 2020, no. 12. Available at: https://www.mdpi.com/2078-2489/11/12/589/htm (Accessed: 11.04.2022).

7. Litvinova T., Sboev A., Panicheva P. Profiling the age of Russian bloggers. *Proceedings of the Conference on Artificial Intelligence and Natural Language*, 2018, pp. 167–177.

8. Khdr A.J., Varol C. Age and Gender Identification by SMS Text Messages. *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, 2018, pp. 1–5.

9. Pentel A. Automatic Age Detection Using Text Readability Features. *EDM*, 2015, no. 1146, pp. 40–45.

10. Madhulika A., Gonçalves T. Age and gender identification using stacking for classification. Teresa. *CLEF*. 2016, pp. 785–790. Available at: https://dspace.uevora.pt/rdpc/handle/10174/20668. (Accessed: 12.04.2022).

11. Argamon S., Koppel M., Pennebaker J., Schler J. Automatically Profiling the Author of an Anonymous Text. *Commun. ACM*, 2009, no. 52, pp. 119–123.

12. Korshunov A., Beloborodov I., Gomzin A. [Determining demographic attributes of microblogging users]. *Proceedings of the Institute for System Programming of the Russian Academy of Sciences*, 2013, no. 25, pp. 179–194.

13. Masi I., Wu Y., Hassner T., Natarajan P. Deep face recognition: A survey. *2018 – the 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2018, pp. 471–478.

14. Devlin J., Chang M.W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv preprint arXiv:1810.04805, 2018. Available at: https://arxiv.org/pdf/1810.04805.pdf (accessed: 13.04.2022).

Artem A. Sobolev
Student, Department of Information System Security
Tomsk State University of Control Systems
and Radioelectronics (TUSUR)
40, Lenin pr., Tomsk, Russia, 634050
Phone: +7-952-183-06-83
Email: bingjo-ya@yandex.ru

Anastasia M. Fedotova
Student, Department of Information System Security
TUSUR
40, Lenin pr., Tomsk, Russia, 634050
Phone: +7-923-444-41-25
Email: afedotowaa@yandex.ru

Anna V. Kurtukova
Postgraduate student, Department of Complex Information
Security of Electronic Computer Systems (KIBEVS), TUSUR
40, Lenin pr., Tomsk, Russia, 634050
Phone: +7-905-991-67-13
Email: av.kurtukova@gmail.com

Aleksandr S. Romanov
Candidate of Science in Engineering, KIBEVS, TUSUR
40, Lenin pr., Tomsk, Russia, 634050
Phone: + 7 (382-2) 41-34-26
Email: alexx.romanov@gmail.com

Alexander A. Shelupanov
Doctor of Science in Engineering, Professor,
Head of Department of Complex Information Security of
Computer Systems, TUSUR
40, Lenin pr., Tomsk, Russia, 634050
Phone: +7 (382-2) 90-71-55, ext. 10-20
Email: saa@fb.tusur.ru