УДК 004.89

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

Применение методов машинного обучения и отбора признаков на основе генетического алгоритма в решении задачи определения автора русскоязычного текста для кибербезопасности

Исследуются подходы к определению автора текста на естественном языке, достоинства и недостатки этих подходов. Процесс идентификации автора русскоязычного текста осуществляется с использованием классических алгоритмов машинного обучения и архитектур нейронных сетей (в том числе fastText, CNN и LSTM и их гибриды, BERT). Оценка эффективности моделей проводится на корпусе сообщений пользователей социальных сетей. Отдельный эксперимент посвящен отбору информативных признаков с помощью генетического алгоритма. Обучение SVM на отобранном генетическим алгоритмом множестве 400 признаков позволяет добиться до 10% прироста точности для всех рассмотренных корпусов авторов. Нейронные сети достигают точности классификации 96%, но при этом их время обучения в некоторых случаях в десятки раз превышает время, затраченное на обучение SVM и других классических методов машинного обучения. Для SVM совместно с генетическим алгоритмом средняя точность составила 66%, для глубоких нейронных сетей и fastText – 73 и 68% соответственно.

Ключевые слова: авторство, анализ текста, машинное обучение, нейронные сети, отбор признаков.
DOI: 10.21293/1818-0442-2021-25-1-79-85

В XXI в. интернет стал коммуникативным пространством информационного общества. У каждого человека появилась возможность высказывать свое мнение и получать отклик от читателей. Множество доступных электронных текстов, в том числе анонимных, указывает на широкий спектр применения методов определения авторства текста [1].

Отдельно стоит отметить публикацию материалов от имени публичных личностей со взломанных аккаунтов. Такой текст способен в кратчайшие сроки стать вирусным и цитируемым, а в случае содержания призывов к запрещенным законом действиям негативно сказаться на настроении в обществе. Сетевые средства массовой информации также активно используются злоумышленниками. Обычно пользователям не требуется предоставлять подлинную информацию о себе – имя, возраст, пол и адрес [2]. Это позволяет анонимно распространять антисоциальную информацию, угрозы и пропаганду терроризма. Методы определения авторства позволяют помочь установить личность создателя текста.

В данном исследовании задача определения автора текста поставлена следующим образом: имеются русскоязычные фрагменты текста, принадлежащие конечному множеству авторов. Авторство некоторых фрагментов установлено. Про остальные анонимные тексты известно, что они принадлежат одному из авторов, но какому – неизвестно. Посредством классификации необходимо определить принадлежность спорных фрагментов истинному автору. В таком случае множеством классов будет являться множество авторов, а тексты, авторы которых известны, – конечной обучающей выборкой. Цель классификации – определить авторство спорных текстов с максимально возможной точностью.

Научная новизна исследования заключается в применении ранее не использовавшихся для русскоязычных текстов методов определения авторства: fastText, комбинации метода опорных векторов (SVM) с генетическим алгоритмом (ГА) для отбора признаков и сравнения этих методов, методы со свёрточной нейронной сетью (CNN), сети с долгой кратковременной памятью (LSTM), их гибриды, представления двунаправленных кодировщиков от Transformers (BERT), *K*-ближайших соседей (KNN), дерево решений (DT), случайный лес (RF) и наивный байесовский классификатор (NB). Следует отметить, что ранее эти методы не использовались для коротких комментариев пользователей социальных сетей.

Существует множество исследований по установлению авторства [3–4]. В ранней работе [1] представлен подробный обзор исследований 2015–2021 гг., включая подходы на основе глубоких нейронных сетей (НС), классических методов машинного обучения (МО), аспектного анализа. В большинстве подобных публикаций применялись различные особенности стиля письма [5], включая лексические, синтаксические, структурные и специфические относительно жанра и тематики текста признаки.

По состоянию на 2022 г. к моделям, успешно решающим смежные задачи текстового анализа, можно отнести LSTM, CNN, их гибриды, fastText, BERT.

При решении многих задач обработки естественного языка немало внимания уделяется качеству векторного представления текста. Созданная в 2016 г. библиотека fastText в реализации от Facebook [6] – серьезный шаг в развитии векторных семантических моделей и методов МО в обработке текста. Преимущество fastText состоит в скорости работы по сравнению с другими моделями. Однако для определения авторства русскоязычных текстов fastText еще не применялся. Поэтому этой модели

решено было уделить особое внимание в данном исследовании.

В большинстве работ применялись различные признаки как в совокупности – в виде их общего вектора, так и по отдельности. Однако не все из них эффективны. Наиболее часто применяют такие характеристики текста, как биграммы и триграммы символов и слов, распределение слов по частям речи, знаки пунктуации. Выделенные признаки могут быть информативными, неинформативными и избыточными. Неинформативные и избыточные признаки бесполезны для классификации. Кроме того, такие признаки могут снизить эффективность классификации из-за большой размерности признакового пространства. Цель отбора признаков [7] – получение подмножества информативных признаков и исключение неинформативных и избыточных. Стоит отметить, что общепринятой комбинации признаков, идентифицирующих автора, не существует, однако биграммы и триграммы символов и слов, распределение слов по частям речи, знаки пунктуации используются в большинстве работ.

Таким образом, целью данного исследования является оценка возможности использования классических методов МО и методов, основанных на НС, для определения автора русскоязычного текста, а также получение подмножества информативных признаков и исключение избыточных и неинформативных для повышения эффективности классификации.

Методы, используемые для определения автора текста

Множеством исследований была доказана эффективность SVM при решении задачи установления авторства текстов [1, 5]. Алгоритмы МО почти всегда требуют структурированных данных, в то время как глубокие НС способны к анализу текстовой последовательности и самостоятельному выделению сетью информативных признаков. Применяются глубокие НС, в частности, такие модели, как LSTM и CNN [8].

В дополнение к стандартным архитектурам глубоких НС, которые были описаны выше, часто используются различные комбинации архитектур [9], например, комбинация нескольких слоев LSTM подряд или CNN с постепенным уменьшением количества фильтров с целью выделить более общие закономерности. Это основано на том, что недостатки одной сети могут компенсироваться преимуществами другой. В данной работе рассмотрены комбинации LSTM и CNN, которые показали отличные результаты в смежной задаче по определению автора исходного кода программы [10]. Стоит отметить, что популярные современные архитектуры CNN с механизмом самовнимания и Transformers в предыдущем исследовании [1] оказались менее точными и наиболее времязатратными, поэтому в этом исследовании более не рассматривались.

Использование простых проверенных методов во многих случаях более оправдано, чем применение новых подходов. Таким образом, в предыдущей работе [2] точность SVM была сопоставима с более современными методами глубокого обучения, в то время как SVM обучался намного быстрее. Поэтому было принято решение расширить список классических методов и протестировать NB, DT, RF, KNN. Преимуществом этих методов является наглядность процесса принятия решения, в отличие от НС, которые представляют собой черный ящик. Результаты классических методов могут быть логически обоснованы, что важно в криминалистике и других областях.

Постановка эксперимента

Важной частью исследования являются сбор и предобработка данных. Модели МО, в частности, глубокие архитектуры, очень требовательны к качеству и объему данных. С этой целью был собран авторский набор данных, включающий большой объем сообщений пользователей социальной сети.

Ещё одним фактором, влияющим на результаты экспериментов, является правильность формирования признакового пространства. В случае с глубокими НС основная сложность состоит уже не в формировании признакового пространства, а в подборе гиперпараметров, управляющих процессом обучения моделей. Даже минимальные изменения этих параметров могут оказать серьезное влияние на результат. Для экспериментальных моделей гиперпараметры подобраны исходя из опыта прошлых исследований авторов [2].

В рамках исследования рассматривалась проблема определения автора текста применительно набору данных, содержащему 202892 коротких комментария 3075 пользователей социальной сети ВКонтакте (VK). Выбор таких данных обоснован максимальной приближенностью к реальным криминалистическим задачам из-за небольшого количества текстов на автора и длины сообщений.

Целью предобработки является очистка набора данных от шумов и избыточной информации, а также преобразование текста в формат, понятный классификатору. В рамках данного исследования подготовка текстов была стандартной:

– перевод всех букв в тексте в нижний регистр;
– удаление стоп-слов;
– удаление цифр и специальных символов;
– форматирование пробельных символов.

На основе обработанного текста строится признаковое пространство текста. При формировании признакового пространства фиксируется совокупность n показателей, измеряемых по каждому тексту. Вектор состоит из различных частотных признаков: частот встречаемости знаков препинания, частей речи, частот униграмм, биграмм и триграмм символов, частот наиболее популярных слов русского языка (основываясь на частотном словаре [11]). Для приведения признаков к общей шкале без потери информации о различии диапазонов применена минимаксная нормализация. Для работы с НС тексты кодируются с помощью метода One-Hot Encoding.

Тексты разделялись в соотношении 80:20 на обучающую и тестовую выборки. Также использо-

валась процедура кроссвалидации. Для оценки качества классификации была рассчитана точность, полученная как доля текстов, по которым классификатор принял правильное решение.

Параметры обучения моделей МО были определены эмпирически, основываясь на опыте предыдущих исследований [12, 13]:

– в качестве алгоритма обучения SVM использовался последовательный метод оптимизации. Линейное ядро. Параметр регуляризации равен 1, а допустимый уровень ошибки задан 0,00001;

– для KNN использовались различные значения *k*, а именно: 3, 5, 7, 15, 25;

– для обучения DT в качестве функции определения качества разбиения использовалась «gini», а максимальная глубина дерева соответствовала 8 разбиениям в ветвях;

– для обучения RF использовалось 5, 15, 25, 35 и 50 деревьев решений.

При обучении глубоких НС в качестве входного слоя каждой сети использовался слой встраивания. Выходной размер слоя – 300, на следующем слое использовался метод исключения с параметром 0,2. Также применялась функция активации выходного слоя – логистическая функция для многомерного случая (Softmax), алгоритм оптимизации – adaptive moment estimation (Adam), метрика – точность. В качестве процедуры оценки эффективности моделей использовалась кроссвалидация по 10 фолдам. Параметры обучения были подобраны экспериментально, исходя из опыта исследователей в области анализа текстовых данных [12]:

– число свёрточных фильтров для CNN – 512, размер ядра – 3. В качестве функции активации была выбрана rectified linear unit (ReLU), а объединяющий слой – GlobalMaxPooling;

– число фильтров для LSTM и BiLSTM – 128. В качестве инструментов прореживания применялись dropout и recurrent dropout с параметром 0,3. Функция активации аналогична CNN;

– в качестве гиперпараметров для гибридов LSTM и CNN использовалось 512 свёрточных фильров, размер ядра – 3, 128 рекуррентных фильтров. Функцией активации выбрана ReLU, объединяющим слоем – GlobalMaxPooling. Процесс прореживания происходил аналогично LSTM. Функция активации аналогична CNN;

– для fastText количество элементов *n*-грамм составило 2–4. Размерность – 50. В качестве функции потерь использована «ova» (softmax loss for multi-label classification). Остальные параметры по умолчанию.

Результаты

В табл. 1 и 2 приведены точности для корпусов из 2, 5, 10, 20 и 50 авторов, а также средняя точность по каждой модели для классических методов МО и НС.

Время обучения всех рассмотренных моделей на корпусе 50 авторов приведено в табл. 3.

Таблица 1

Результаты определения автора с использованием МО, обученных на признаковом пространстве

	Точность моделей, %					
	2	5	10	20	50	Средняя точность
SVM	**72±4**	**70±4**	**66±4**	**65±3**	32±4	**59±3**
NB	63±2	59±3	46±4	39±4	29±3	47±3
DT	69±2	54±2	34±4	30±4	26±3	44±3
RF	71±4	56±3	38±3	32±2	26±2	46±3
KNN	68±4	65±4	62±4	44±4	**34±3**	55±4

Таблица 2

Результаты определения автора с использованием НС

Модели	Точность моделей, %					
	2	5	10	20	50	Средняя точность
LSTM	93±2	90±2	73±3	69±2	50±3	75±2
BiLSTM	95±2	**93±2**	71±2	59±1	50±4	74±2
CNN	**96±2**	**93±2**	72±3	**68±1**	49±4	**76±3**
CNN+ LSTM	**96±4**	91±2	**77±4**	62±3	47±2	75±3
LSTM+ CNN	92±2	90±1	64±3	61±1	47±3	71±3
fastText	94±1	87±2	76±4	**68±2**	**56±3**	**76±2**
RuBERT	93±2	89±2	**77±3**	67±3	50±3	75±3
Multi-BERT	90±2	87±2	70±3	63±3	47±3	72±3

Таблица 3

Время обучения моделей на корпусе 50 авторов

Модели НС	Время обучения, с	Модели МО	Время обучения, с
LSTM	30190	SVM	589
BiLSTM	32980		
CNN	25380	NB	308
CNN+LSTM	26467		
		DT	236
LSTM+CNN	28397		
		RF	804
fastText	15926		
RuBERT	26547	KNN	644
MultiBERT	27117		

Полученные результаты позволяют сделать вывод о неэффективности классических методов МО даже при использовании сформированного признакового пространства. Это связано с тем, что длины комментариев очень малы и в среднем составляют всего 13,3 символа. Содержимое комментариев отражает эмоции автора по отношению к комментируемому событию, поэтому преобладают односложные высказывания. В связи с этим объем текста слишком мал для получения индивидуальных характеристик автора даже на тщательно сформированном пространстве признаков. SVM, обученный на экспериментально подобранных параметрах и признаковом пространстве, достигает максимальной точности 72% для двух авторов, в то время как глубокие НС способны классифицировать с точностью 96% при той же сложности задачи. Данный факт объясняется способностью глубоких НС к самостоятельному выделению неявных информативных при-

знаков. FastText превосходит по точности LSTM+CNN для всех рассмотренных наборов авторов и обучается в среднем на 39% быстрее. Также для 2 и 10 авторов точность fastText выше, чем BiLSTM, а в случае 50 авторов превосходит все остальные модели. По скорости обучения fastText превосходит все рассмотренные глубокие НС в среднем на 42%.

Отбор информативных признаков с помощью генетического алгоритма

Отбор признаков на основе генетического алгоритма позволяет выделить оптимальное подмножество из общего множества признаков. Помимо прироста в точности за счет удаления избыточных и неинформативных признаков, такое решение позволяет ускорить обучение модели.

Всего в ГА используются три оператора: селекция, скрещивание, мутация. Для работы алгоритма необходимо задать вероятности для операторов скрещивания и мутации, а также задать тип селекции. С увеличением количества поколений возрастает вероятность нахождения глобального оптимума – популяции, гены которой являются наиболее приспособленными. Критерием остановки могут служить пороговое значение точности классификации, исчерпание времени работы алгоритма или заданного числа обращений к целевой функции, определенной как средняя точность SVM по кроссвалидации. Хромосома представляет собой бинарный вектор признаков, где сами признаки представляют собой гены, 1 соответствует вхождению признака в множество информативных, 0 означает исключение признака из оптимального множества.

В данном исследовании ГА использовался совместно с SVM. Такой выбор обоснован тем, что именно SVM в большинстве случаев продемонстрировал лучшую точность среди классических методов МО. В применении ГА совместно с глубокими НС нет необходимости ввиду способности таких архитектур к самостоятельному поиску информативных признаков. Работа алгоритма выполнялась до достижения заданного количества итераций, затем выбиралось лучшее решение – то, где значение целевой функции максимально. В качестве типа селекции была выбрана элитная селекция. Суть выбранного типа в том, что выбираются лучшие признаки на основе сравнения значений точности. Далее они вступают в различные преобразования, после которых выбираются новые элитные элементы, данный процесс продолжается до момента прекращения появления элитных элементов.

Одной из сложностей использования ГА является задание значений вероятностей для операций мутации и скрещивания. В настоящее время нет общепринятых норм и правил, согласно которым требуется выбрать конкретные значения, поэтому в данном исследовании значения параметров операций скрещивания и мутации подбирались экспериментально SVM. ГА задан следующими параметрами:

– коэффициент скрещивания: 0,5;
– коэффициент мутации: 0,2;
– количество популяций: 1000.

Эксперименты проводились для получения 50, 100, 200, 300, 400, 500 признаков из исходных 1168. При полном переборе время отбора таких подмножеств очень велико, поэтому использовалось ограничение по количеству обращений к целевой функции. Для получения 50, 100, 200, 300, 400 и 500 признаков в результирующем наборе было установлено ограничение на выбранное количество признаков как максимальное. Результаты экспериментов представлены в табл. 4.

Таблица 4

Результаты отбора признаков для комментариев пользователей социальной сети

Кол-во признаков	Точность моделей, %					
	2	5	10	20	50	Средняя точность
50	65±6	50±4	47±2	44±4	22±4	46±4
100	65±4	59±4	57±4	49±4	27±3	52±4
200	67±5	64±4	60±4	53±4	27±4	54±4
300	75±5	72±2	67±3	59±3	34±3	61±3
400	**80±3**	**77±3**	**72±3**	**65±3**	**37±4**	**66±3**
500	75±3	70±3	68±4	62±3	35±4	62±4
1168	72±4	70±4	66±4	**65±3**	32±4	59±3

Исходя из представленных результатов, уменьшение количества признаков более чем вдвое не только не снижает точность классификации, но и позволяет улучшить результат определения автора текста. Точность, полученная для 200 признаков, сопоставима с исходной. 100 и 50 признаков не являются достаточными для определения автора. Набор 400 признаков, на котором была достигнута максимальная точность, содержит частотные распределения 6 знаков пунктуации, 8 частей речи, 20 униграмм символов, 107 биграмм символов, 98 триграмм символов и 165 слов из частотного словаря.

Чтобы проверить наличие статистически значимой разницы между результатами, полученными SVM, обученным с использованием различного количества информативных признаков, выбранных ГА, и различными результатами перекрестной проверки, были применены ранговые апостериорные тесты Фридмана [14] и Неменьи [15]. Тесты проводились для самого сложного из рассмотренных случаев – классификации 50 авторов. Нулевая гипотеза заключалась в том, что разница между результатами, полученными на разном количестве признаков, является только случайной. Альтернативная гипотеза заключалась в том, что существует статистически значимая разница между результатами. Значение p составило 0,0007. Поскольку это значение меньше 0,05, нулевую гипотезу можно отвергнуть.

Эффективность методов существенно различается, если соответствующие средние ранги отличаются хотя бы на величину критической разницы. Чтобы оценить разницу, был применен апостериорный тест Неменьи после отклонения нулевой гипотезы теста Фридмана. Данный тест Неменьи предна-

значен для обнаружения различных групп данных. Результаты представлены на рис. 1 в виде диаграммы значимости Демшара [16]. В случае, если разница средних рангов между двумя методами меньше, чем автоматически рассчитанное значение критической разницы, разница в их производительности незначительна и на рисунке представлена горизонтальной линией.

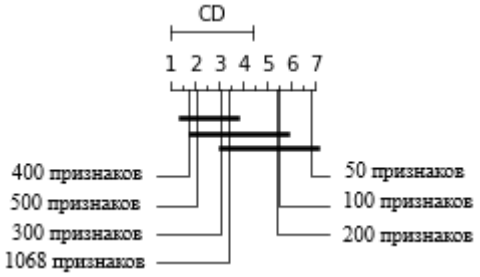

Рис. 1. Диаграмма значимости Демшара

Как видно из диаграммы, уменьшение до 50–100 информативных признаков отрицательно сказалось на точности. Тот же результат был достигнут с использованием всего пространства 1168 признаков. Наборы из 200–400 признаков позволили добиться сопоставимой точности классификации. Лучшим вариантом признано использование SVM, обученного на 400 признаках, полученных ГА, худшим – ограничение количества признаков до 50.

Заключение

Реализованные в работе методы показывают результаты, сопоставимые и превосходящие рассмотренные в рамках обзора аналоги. Для классических методов МО рассмотрена классификация на сформированном признаковом пространстве. В данном случае не удаётся достичь результатов, сопоставимых с полученными при обучении глубоких НС. Для различного количества авторов разница в точности варьируется от 2 до 30%. Обоснование этого факта кроется в недостаточной для формирования вектора длине односложных высказываний и предложений, математически описывающего авторский инвариант.

С целью улучшения качества классификации SVM проведен отбор информативных признаков ГА. В рамках отбора поставлена задача максимизации целевой функции. Из исходного множества 1168 признаков отбирались подмножества 500, 400, 300, 200, 100 и 50 признаков согласно значению целевой функции. Такое решение позволяет не только выделить информативные признаки, но и устранить избыточные, затрудняющие классификацию. Векторы, состоящие из 50 признаков, не позволяют улучшить классификацию. Обучение SVM на отобранном ГА множестве 400 признаков позволяет добиться до 10% прироста точности для всех рассмотренных корпусов авторов. Подобное решение позволяет ускорить процесс обучения классификатора, снизить нагрузку на вычислительные ресурсы и устранить избыточность набора признаков.

Глубокие НС, в отличие от SVM, способны самостоятельно выявлять неявные информативные признаки для классификации. При обучении CNN получена точность 96%, что превосходит точность SVM, обученном на отобранном множестве признаков, и является максимальным результатом во всей серии экспериментов. Точность сетей с долгой кратковременной памятью, в том числе двунаправленных, а также их комбинации с CNN достигают высокой точности для всех наборов данных, но при этом их время обучения в некоторых случаях в десятки раз превышает время, затраченное на обучение SVM и других классических методов МО. Оптимальным вариантом является fastText, скорость обучения которого в среднем на 51% меньше, чем для других рассмотренных глубоких НС, а точность ниже максимальной по всем моделям не более чем на 3%.

При выборе метода определения автора текста следует руководствоваться оценкой подлежащих исследованию текстов, учитывая их объем, количество образцов и источник. В случае ограниченности ресурсов следует использовать классические методы МО и ГА. В случае возможности умышленного искажения текста или вероятности атаки больше подойдут глубокие НС ввиду способности выявления неявных признаков авторского стиля.

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. Романов А.С. Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста / А.С. Романов, А.А. Шелупанов, Р.В. Мещеряков. – Томск: В-Спектр, 2011. – 188 с.

2. 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: 25.12.2021).

3. Boenninghoff B. Deep bayes factor scoring for authorship verification // arXiv preprint arXiv:2008.10105. – 2020. – URL: https://arxiv.org/abs/2008.10105, свободный (дата обращения: 26.12.2021).

4. Jafariakinabad F. Self-supervised Representation Learning of Sentence Structure for Authorship Attribution / F. Jafariakinabad, K.A. Hua. – arXiv preprint arXiv:2010.06786. 2020. – URL: https://arxiv.org/pdf/2010.06786.pdf, свободный (дата обращения: 26.12.2021).

5. Uchendu A. Authorship Attribution for Neural Text Generation // Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2020. – P. 8384–8395.

6. FastText: Library for efficient text classification and representation learning. – URL: https://fasttext.cc, свободный (дата обращения: 28.12.2021).

7. Sarin K.S. Bagged ensemble of fuzzy classifiers and feature selection for handwritten signature verification / K.S. Sarin, I.A. Hodashinsky. – Computer Optics. – 2019. – Vol. 43, No. 5. – P. 833–845.

8. Explainable Authorship Verification in Social Media via Attention-based Similarity Learning / B. Boenninghoff, S. Hessler, D. Kolossa, M. Nickel // IEEE International Conference on Big Data (Big Data). – 2019. – P. 36–45. – URL: https://arxiv.org/pdf/1910.08144, свободный (дата обращения: 30.12.2021).

9. Исхакова А.О. Метод и программное средство определения искусственно созданных текстов: дис. … канд. техн. наук. – Томск: ТУСУР, 2016. – 123 с.

10. 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: 04.01.2022).

11. Ляшевская О.Н., Шаров С.А. Новый частотный словарь русской лексики. – URL: http://dict.ruslang.ru/freq.php (дата обращения: 06.01.2022).

12. 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: 07.01.2022).

13. Natural Text Anonymization Using Universal Transformer with a Self-attention / A. Romanov, A. Kurtukova, A. Fedotova, R. Meshcheryakov // Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL – 2019), November 27, 2019. – Saint-Petersburg, Russia, 2019. – P. 22–37.

14. Friedman Test in SPSS Statistics. – URL: https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php (дата обращения: 09.01.2022).

15. Friedman Test Post-hoc Analysis. – URL: https://www.real-statistics.com/anova-repeated-measures/friedman-test/friedman-test-post-hoc-analysis (дата обращения: 09.01.2022).

16. CD diagrams for the post-hoc Nemenyi test. – URL: https://www.imsbio.co.jp/RGM/R_rdfile?f=performanceEstimation/man/CDdiagram.Nemenyi.Rd&d=R_CC (дата обращения: 09.01.2022).

Куртукова Анна Владимировна
Аспирант каф. комплексной информационной безопасности электронно-вычислительных систем (КИБЭВС) Томского государственного университета систем управления и радиоэлектроники (ТУСУР)
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

Федотова Анастасия Михайловна
Студентка каф. безопасности информационных систем (БИС) ТУСУРа
Ленина пр-т, 40, г. Томск, Россия, 634050
Тел.: +7-923-444-41-25
Эл. почта: fedotova.a.747@e.tusur.ru

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

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

Application of machine learning methods and feature selection based on a genetic algorithm in solving the problem of determining the authorship of a Russian-language text for cybersecurity

The article explores the approaches to determine the author of a natural language text, the advantages and disadvantages of these approaches. The identification is carried out using classical machine learning algorithms and neural network architectures (including fastText, CNN and LSTM and their hybrids, BERT). The efficiency of the model is evaluated based on the social media texts dataset. A separate experiment is devoted to the feature selection using a genetic algorithm. SVM trained on a selected 400 features set makes it possible to achieve up to 10% increase in accuracy for all considered numbers of authors. Neural networks achieve a classification accuracy of 96%, but their training time in some cases exceeds the time spent on training SVM and other classical machine learning methods in some cases. For SVM together with the genetic algorithm, the average accuracy was 66%, for deep neural networks and fastText – 73 and 68%, respectively.

Keywords: authorship, text mining, machine learning, neural networks, deep learning, feature selection.
DOI: 10.21293/1818-0442-2021-25-1-79-85

References

1. Romanov A.S., Shelupanov A.A., Meshcheryakov, R.V. [Development and Research of Mathematical Models, Methods and Software Tools of Information Processes in the Identification of the Author of the Text]. Tomsk, V-Spektr, 2011, 188 p. (in Russ.).

2. Romanov A., Kurtukova A., Shelupanov A., Fedotova A., Goncharov V. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Net-works. *MDPI Future Internet*, 2021, no. 1. Available at: https://www.mdpi.com/1999-5903/13/1/3/htm (Accessed: December 25, 2021).

3. Boenninghoff B. Deep bayes factor scoring for authorship verification. arXiv preprint arXiv:2008.10105. 2020. Available at: https://arxiv.org/abs/2008.10105 (Accessed: December 26, 2021).

4. A Self-supervised Representation Learning of Sentence Structure for Authorship Attribution. arXiv preprint arXiv:2010.06786. 2020. Available at: https://arxiv.org/abs/2010.06786 (Accessed: December 26, 2021).

5. Uchendu A. Authorship Attribution for Neural Text Generation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8384–8395.

6. FastText: Library for efficient text classification and representation learning. Available at: https://fasttext.cc (Accessed: December 28, 2021).

7. Sarin K.S., Hodashinsky I.A. Bagged ensemble of fuzzy classifiers and feature selection for handwritten signature verification. *Computer Optics*, 2019, vol. 43, no. 5, pp. 833–845.

8. Boenninghoff B., Hessler S., Kolossa D., Nickel M. Explainable Authorship Verification in Social Media via At-

tention-based Similarity Learning. *Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019.* Available at: https://arxiv.org/pdf/1910.08144 (Accessed: December 30, 2021).

9. Iskhakova A.O. [Method and Software for Determining Artificially Created Texts]. Cand. Diss. *Proceedings of TUSUR University,* 2016, 123 p. (in Russ.).

10. 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: January 4, 2022).

11. Lyashevskaya O. N., Sharov S. A. New frequency dictionary of Russian vocabulary. Available at: http://dict.ruslang.ru/freq.php (Accessed: January 6, 2022) (in Russ.).

12. 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: January 7, 2022).

13. Romanov A., Kurtukova A., Fedotova A., Meshcheryakov R. Natural Text Anonymization Using Universal Transformer with a Self-attention. *Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL – 2019)*. Saint Petersburg, Russia, November 27, 2019, pp. 22–37.

14. Friedman Test in SPSS Statistics. Available at: https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php (Accessed: January 9, 2022).

15. Friedman Test Post-hoc Analysis. Available at: https://www.real-statistics.com/anova-repeated-measures/friedman-test/friedman-test-post-hoc-analysis (Accessed: January 9, 2022).

16. CD diagrams for the post-hoc Nemenyi test. Available at: https://www.imsbio.co.jp/RGM/R_rdfile?f=performanceEstimation/man/Cddiagram.Nemenyi.Rd&d=R_CC (Accessed: January 9, 2022).

Anna V. Kurtukova
Postgraduate student, Department of Complex Information Security of Electronic Computer Systems (KIBEVS),
Tomsk State University of Control Systems and Radioelectronics (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, Associate professor, Department of KIBEVS, TUSUR
40, Lenin pr., Tomsk, Russia, 634050
Phone: +7 (382-2) 41-34-26
Email: alexx.romanov@gmail.com

Anastasia M. Fedotova
Student, Department of Information System Security (ACS), TUSUR
40, Lenin pr., Tomsk, Russia, 634050
Phone: +7-923-444-41-25
Email: fedotova.a.747@e.tusur.ru

Alexandr A. Shelupanov
Doctor of Science in Engineering, Professor,
President TUSUR
40, Lenin pr., Tomsk, Russia, 634050
Phone: +7 (382-2) 90-71-55
Email: saa@tusur.ru