МЕТОДОЛОГИЯ ИДЕНТИФИКАЦИИ АВТОРА ТЕКСТА ДЛЯ РЕШЕНИЯ ЗАДАЧ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Романов А. С.[1]

DOI: 10.21681/2311-3456-2024-3-120-128

Цель работы: *создание методологии идентификации автора текстовой информации, включая естественно-языковые тексты и исходные коды программ, для решения задач информационной безопасности.*

Объектом исследования *является печатный текст и его характеристики.*

Предметом исследования *являются характеристики текста, описывающие авторский стиль, методы и алгоритмы машинного обучения, предназначенные для работы с естественно- и искусственно-языковыми текстами.*

Методы исследования *включают методы теории множеств, математической статистики, вычислительного эксперимента и методы искусственного интеллекта.*

Научная новизна: *предложена комплексная методология идентификации автора текста, учитывающая особенности естественно- и искусственно-языковых текстов, а также предложена модель создания текста автором в киберсреде, впервые учитывающая семантические особенности и информативные признаки текста на разных уровнях иерархического анализа, специфику среды, атрибуты автора и вид деятельности по созданию текста.*

По результатам исследования *предложена методология идентификации автора естественно-языкового текста и исходных текстов программ для решения задач информационной безопасности в виде комплекса методов, моделей и алгоритмов, агрегирующий имеющийся опыт. Методология является универсальной для решения задач информационной безопасности, связанных с классификацией текстов.*

Ключевые слова: *интеллектуальный анализ текста, семантика, машинное обучение, исходный код, атрибуция.*

METHODOLOGY FOR IDENTIFYING THE AUTHOR OF TEXT INFORMATION FOR SOLVING CYBERSECURITY TASKS

Romanov A. S.[2]

The goal of article: *the creation of a methodology for identifying the author of textual information, including natural language texts and program source codes, is aimed at solving information security issues.*

The object of study: *printed text and its characteristics.*

The subject of study: *characteristics of text that describe the author's style, methods, and machine learning algorithms designed for processing both natural and artificially-generated texts.*

The research methods: *set theory methods, mathematical statistics, computational experiments, and methods of artificial intelligence*

Scientific novelty: *for the first time, a comprehensive methodology for identification of a text's author and a model for text creation by an author in a cyber environment have been proposed. The proposed methodology considers features of both natural and artificially-generated texts. An introduced model takes into account semantic features and informative characteristics of the text at different levels of hierarchical analysis, specifics of the environment, author attributes, and the nature of activities involved in creating the text.*

1 Романов Александр Сергеевич, кандидат технических наук, доцент, доцент кафедры Комплексной информационной безопасности электронно-вычислительных систем (КИБЭВС), Томский государственный университет систем управления и радиоэлектроники, Томск, Россия. ORCID: 0000-0002-2587-2222. Scopus Author ID: 57221288963. E-mail: alexx.romanov@gmail.com

2 Romanov Aleksandr Sergeevich, Candidate of Technical Sciences, Associate Professor, Department of Complex Information Security of Electronic Computing Systems, Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia. ORCID: 0000-0002-2587-2222. Scopus Author ID: 57221288963. E-mail: alexx.romanov@gmail.com

Results obtained: *a methodology has been proposed for identifying the author of a natural language text and program source codes to address information security challenges. The methodology includes a set of methods, models and algorithms that aggregate existing research experience. The methodology is universal for solving information security issues related to text classification.*

Keywords: *text mining, semantics, machine learning, source code, attribution.*

Введение

Один из видов нарушений в киберпространстве – нарушение авторских и смежных прав на текстовые произведения, которое может выражаться, например, в присвоении текста другого человека с целью получения материальной выгоды или попытке выдать авторство созданного текста за авторство другого лица. Методы определение авторства позволяют выявить подобные нарушения и установить личность создателя текста. Атрибуция текстов является важной проблемой в компьютерной лингвистике, журналистике, а также в криминалистике, где знание истинного автора анонимного текста (например, предсмертной записки) может облегчить и ускорить работу правоохранительных органов. Таким образом проблема идентификации автора текста для защиты интеллектуальной собственности является важной задачей информационной безопасности.

В числе атрибутов автора выделяют пол, возраст, образование, профессию, личностные качества и др. Первичными считают гендерный признак и возраст, потому что их конкретизация позволяет установить остальные атрибуты (вторичные) и сузить круг кандидатов при определении автора. Совокупность атрибутов формирует уникальную языковую личность. Одним из направлений практического применения методик определения пола и возраста является криминалистика, где важно определение психологического портрета преступника и профилирование автора. В контексте проблем информационной безопасности подобные методики являются основой для мониторинга социальных сетей для выявления информации, пропагандирующей нетрадиционные сексуальные отношения, педофилию, смену пола. Важной проблемой также является недопущение детей и подростков до запрещенного или шокирующего контента, который имеет возрастное ограничение «18+», либо ограничение общения с целью предотвращения педофилии. Определение гендерной принадлежности автора текста актуально поскольку решением Верховного Суда Российской Федерации от 30.11.2023 по делу № АКПИ23-990С «движение ЛГБТ» признано экстремистской организацией и его деятельность запрещена на территории России. Таким образом определение пола и возраста по тексту для дальнейшего выявления признаков пропаганды ЛГБТ и педофилии является важной задачей информационной безопасности, актуальным является также вопрос создания программного обеспечения для мониторинга социальных сетей.

Методы определения однородности авторского стиля сообщений можно использовать для продленной аутентификации [1] в социальной сетях и мессенджерах, обнаружения аномалий и необычных паттернов в потоке текстовых данных пользователей сети Интернет. Таким образом задача идентификации автора сообщений в сети Интернет и продленная аутентификация пользователя социальных сетей на основе текста является важной задачей информационной безопасности, имеющей существенные особенности в методологическом плане.

Анализ и определение авторства текста с учетом эмоциональной составляющей имеют особую важность в контексте борьбы с экстремизмом и терроризмом [2–5]. В современном информационном обществе социальные сети и онлайн-платформы стали пространством для распространения экстремистских и радикальных идей. Опасность заключается в том, что Интернет-платформы предоставляют экстремистам доступ к широкой аудитории и возможность быстрой и масштабной пропаганды своих идей. Это может воздействовать на уязвимых или подверженных влиянию лиц, включая молодежь, подстрекая их к насилию, террористическим действиям или участию в экстремистских организациях. Мониторинг этой информации в определенные моменты времени и выявление лиц, имеющих целью совершение злонамеренных действий, становится актуальной практической задачей противостояния террористической угрозе и защиты государства. Таким образом анализ настроения автора, определения эмоциональной окраски текста, деструктивных, а также текстов экстремистской направленности, запрещенных законодательством Российской Федерации, является важной задачей информационной безопасности.

Решение задачи определения автора программного кода являются критически значимыми для обеспечения безопасности в цифровой среде. Это связано с тем, что подавляющее большинство технологий, а также программных систем и комплексов, упрощающих профессиональную и повседневную деятельность человека, подвержены

сбоям. Подобные проблемы могут возникать по ряду причин, например, в результате ошибок разработчиков (при проектировании, реализации и/или внедрении), неправильной эксплуатации пользователями, неполадок смежных систем. Однако наибольшую угрозу несут сбои, происходящие ввиду преднамеренного и/или злоумышленного вмешательства. Несмотря на то, что деятельность по созданию, использованию и распространению вредоносного программного обеспечения запрещена на законодательном уровне и закреплена в ст. 272, 273, 274 Уголовного Кодекса Российской Федерации, технические средства, обеспечивающие эффективное и своевременное установления авторства исходного кода в рамках компьютерных экспертиз, на 2024 год отсутствуют. Анализ исходных кодов программ на предмет авторства осуществляется специалистами в области компьютерной криминалистики вручную или с использованием малоэффективных для данной задачи средств текстового анализа. Таким образом, задача определения автора-вирусописателя представляют особую важность для информационной безопасности.

Актуальной задачей информационной безопасности становится проблема создания и усовершенствования методик, учитывающих способы сокрытия авторского стиля (обфускация) и имитации авторского стиля [6], а также генеративных моделей, позволяющих автоматически генерировать тексты на основе глубоких моделей, обученных на больших текстовых корпусах (GPT). В связи с этим возникает необходимость в проведении дополнительных исследований, направленных на оценку устойчивости методов определения авторства текста к такого рода атакам.

Анализ предметной области

В настоящее время наблюдается повышенный интерес к количественным методам анализа текстовой информации на основе слабо контролируемых человеком характеристик текста, общих для всех авторов. С развитием методов текстового анализа, в работах по определению авторства начинает превалировать использование семантической информации о тексте [7], наряду с лексическими, морфологическими и синтаксическими признаками.

Диссертационная работа Москина Н. Д. [8] и связанные с ней исследования [9] фокусируются на разработке и модернизации теоретико-графовых моделей, использующих технологию Graph Neural Network (GNN), для определения авторства текстов. В исследовании анализируются 500 текстов неопределенного авторства и более 800 произведений русских классиков. Основные методы включают агрегацию графов, учет их иерархичности, нечеткости и темпоральности, а также использование метрик, таких как максимальный общий подграф. Проверялись гипотезы о различиях в структурных характеристиках графов разных жанров. Ансамбль моделей был разработан для описания языковой структуры текстов. Для анализа использовались такие методы, как рекуррентная НС, сеть долгой краткосрочной памяти (LSTM), Transformer, дерево решений, SVM, деревья решений (RF). Наилучший результат показала модель Transformer с точностью 97%, в то время как дерево решений показало минимальную эффективность в 43%.

В диссертационной работе Огорелкова И. В. [10] исследуются гендерные различия в русскоязычных политических текстах. Общий корпус текстов состоит из 1000 произведений, разделенных на мужской и женский корпусы по 500 текстов каждый. Исследование включает анализ четырех основных групп признаков: смысловых, текстологических, языковых и психолингвистических, дополненных лексическими и синтаксическими особенностями. В работе выделено 20 ключевых признаков [11], специфичных для мужской и женской письменной речи, такие как использование определенных союзов, частиц, местоимений и вводных слов. Эти признаки затем анализируются для определения пола автора текста. Заключительный этап включает оценку информативности каждого признака и окончательное определение гендерной принадлежности автора. Выводы исследования указывают на характерные различия в стиле мужской и женской речи, такие как лаконичность и аргументированность для мужчин, против многословия и эмоциональности для женщин.

В диссертации Сбоева А. Г. [12] и соответствующей статье [13] обсуждается методика определения пола и возраста автора русскоязычного текста на основе морфологических, синтаксических признаков, n-грамм, токенов, частей речи, эмоциональных признаков и эмбеддингов. Использованы методы глубокого обучения, включая сиамскую нейронную сеть и оригинальную архитектуру SyntGraphLSTM, с моделью представления текста TF-IDF и классификаторами SVM и RF. Корпус состоял из 1850 контролируемых и 41624 реальных текстов из социальных сетей, включая 4332 текста с искаженным полом и 13632 с искаженным возрастом. Методика показала точность в 86% по метрике F1 для определения пола, 64% при намеренных искажениях, 48% для определения возрастной группы (выше случайного угадывания на 15%) и 44% для распознавания искаженного возраста, при этом направление искажения определялось с точностью 80%. Возрастные группы включали 18–23, 24–29 и старше 30 лет.

В диссертации Давыдовой Ю. В. [14] исследуются методы мониторинга контента в социальных сетях с целью реализации превентивных мер пропаганды криминализации. Методы включали технику специализированного текстового поиска на основе динамического программирования и деревьев

решений, новый подход, основанный на семантическом анализе для определения жаргонизмов и неологизмов, тематической лексики пропагандистов криминала, моделирование ошибок на основе гибридной модели, сочетающей лингвистические правила и статистические данные для присвоения веса различным типам текстовых ошибок. Данные для исследования включали федеральные и региональные базы данных правоохранительных органов и платформы социальных сетей. Был разработан прототип программного обеспечения, который апробирован в реальных сценариях для мониторинга и анализа социальных сетей на предмет незаконной деятельности. Предложенный подход продемонстрировал точность 95% при обнаружении противозаконной информации.

В диссертационной работе Андреева И. А. [15] предложена методика построения социального портрета пользователя в рамках подбора кадров с учетом материальных. профессиональных и социальных рисков работодателя. Предложен подход к унификации и агрегации данных из различных социальных сетей, сопоставления профилей пользователей разных социальных сетей, обработка слабоструктурированных данных. Целью работы является формирование психоэмоционального портрета пользователя на основе тонального анализа созданных им текстов и информации из профилей социальных сетей. Набор данных включал более 2,5 миллионов сообщений. Классификация осуществлялась на основе категоризации текстов по 10 эмоциям. При формировании вектора признаков был расширен словарь WordNet-Affect, отдельно проводился анализ эмотиконов. Для проведения экспериментов были отобраны 100 пользователей, имеющих 1 и более аккаунт в социальных сетях. Максимальная точность 87% была достигнута моделью BERT, классические методы машинного обучения (SVM LR, RF) оказались менее эффективны, при их использовании точность составляла не более 65% при использовании SVM.

В диссертации Стремоухова В. Д.[3] рассматривается задача определения авторства бинарного кода, особенно вредоносных программ. Исследование включает применение математического анализа и статистики, таких как Марковские процессы и корреляционный анализ. Стремоухов предлагает модели, в том числе на основе сжатия данных и относительной энтропии, которые анализируют объединение анонимного кода с известными образцами для оценки степени сжатия. Другая модель использует матрицу переходных вероятностей Маркова и способна отсеивать неинформативные части кода. Модели апробированы на исполняемых файлах win32 и вирусных коллекциях «Лаборатории Касперского», показав эффективность с точностью до 100%. Однако, метод требует единообразия в языке программирования, платформе и компиляторе, что ограничивает его применение в сложных случаях.

Перечисленные исследователи успешно решали частные задачи атрибуции, не связанные с обеспечением информационной безопасности напрямую. Однако проблема систематизации существующих подходов к идентификации автора текста, а также разработки комплексной методологии, позволяющей эффективно решать взаимосвязанные задачи авторства в интересах национальной безопасности страны, остается открытой.

Методология идентификации автора текстовой информации

На основе проведенного анализа разработана обобщенная методология идентификации автора текстовой информации для решения задач информационной безопасности (рис. 1).

Представленная на рис. 1 методология оперирует множеством взаимосвязанных моделей, методов и алгоритмов для анализа естественно- и искусственно-языковых текстов и включает:

1. Модель создания автором текста в киберсреде. Ключевая модель методологии, описывающая процесс создания текстов авторами с учетом особенностей и ограничений среды.

2. Модели представления текста. Тексты, подлежащие анализу, содержат множество явных и неявных признаков, указывающих на различные авторские характеристики: пол, возраст, опыт, идеологию, настроение и др. Каждый такой признак может оказать существенное влияние на конечный результат, поэтому важно представить текст в виде информативных признаков.

3. Алгоритмы разбора текста. Для разбора текста на различных уровнях (лексическом, морфологическом синтаксическом, семантическом) в рамках текстового анализа используются различные методы и инструменты машинного обучения и обработки естественного языка.

4. Методы принятия решений. Ключевой частью методологии является процесс принятия решения о принадлежности текста к классу в зависимости от задачи. Решение принимается статистическими методами, использующими меры расстояния и сходства признаков в пространстве, методами на основе машинного обучения, методами на основе глубокого обучения. Каждая группа методов имеет свои преимущества и недостатки по отношению к конкретной задаче текстового анализа. При этом методы могут применяться как по отдельности, так и совместно в качестве алгоритмического ансамбля.

5. Методы оптимизации гиперпараметров. Большинство задач машинного обучения сводятся к поиску параметров модели, которые минимизируют некоторую функцию потерь. Функция потерь оценивает, насколько хорошо модель соответствует данным.

3 Стремоухов В. Д. Модель и метод анализа схожести и определения авторства вредоносного кода: дис. канд. техн. наук: 05.13.19 / В. Д. Стремоухов. – НИУ ИТМО, Санкт-Петербург, 2013. – 95 с.

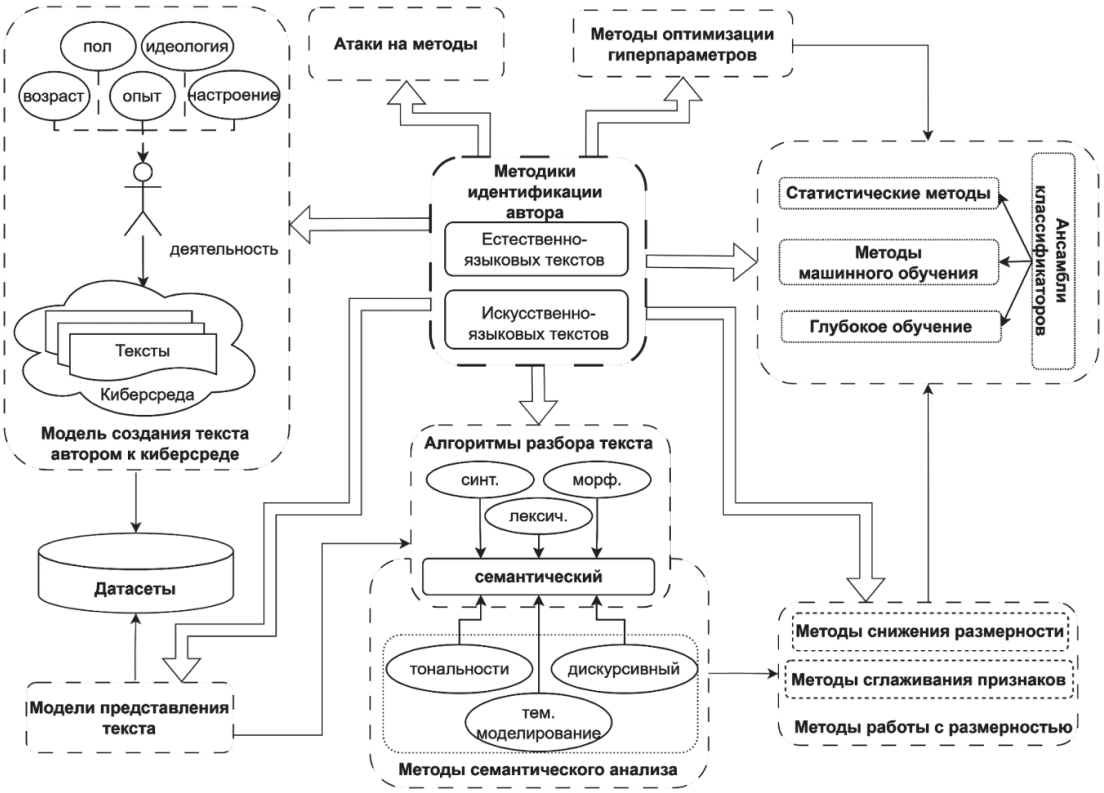

Рис. 1. Методология идентификации автора текстовой информации для решения задач информационной безопасности

Благодаря оптимизации гиперпараметров, обучение становится более эффективным и быстрым, обеспечивается сходимость алгоритма и повышается качество конечной модели.

6. Методы снижения размерности и отбора признаков. Этот процесс помогает повысить эффективность и качество моделей машинного обучения. Во-первых, за счет снижения размерности уменьшается вычислительная сложность. Модели машинного обучения, особенно полученные путем глубокого обучения, могут быть требовательны к вычислительным ресурсам. Уменьшение числа признаков может существенно сократить время обучения и предсказания, а также снизить требования к памяти. Во-вторых, применение информативных признаков в противовес полному множеству признаков позволяет улучшить обобщающую способность модели и предотвратить проблему переобучения за счет удаления нерелевантных или избыточных признаков. В-третьих, в результате снижения размерности признакового пространства повышается интерпретируемость модели: модели с меньшим числом признаков легче интерпретировать.

7. Атаки на методы. Атаки на методы – это специальные методы, предназначенные для запутывания моделей машинного обучения. Среди таких атак можно выделить обфускацию текста, искусственную-генерацию текста и усреднение признаков. Обфускация подразумевает добавление шума в виде опечаток, замену слов синонимами или бессмысленными словами, изменение порядка слов, применение гомоглифов (символов иноязычного алфавита, визуально похожих на исходный), а также использование перефразирования. Искусственная генерация позволяет выполнять имитацию стиля автора за счет использования современных генеративных моделей семейства GPT. С их помощью создаются реалистичные тексты, способны запутать классификатор. Эти атаки могут быть статическими, созданными один раз и используемыми без изменений, или динамическими, адаптирующимися к обновлениям модели классификации. Применяемые для решения задач информационной безопасности модели классификации текстов должны быть устойчивыми к атакам на метод.

8. Алгоритмы сглаживания. Информативные признаки текста, свойственные текстам больших объемов, могут не проявиться в текстах небольших объемов. Алгоритмы сглаживания позволяют оценить вероятности не наступивших событий.

9. Методики текстового анализа. Методология идентификации автора текстовой информации является основой для создания частных методик, позволяющих решать практические задачи информационной безопасности. Методика анализа естественно-языкового текста [16] позволяет решать задачи открытой и закрытой атрибуции автора текста. Закрытая атрибуция является более простым

случаем, т.к. подразумевает наличие истинного автора текста среди авторов-кандидатов. Открытая атрибуция является более сложным случаем ввиду отсутствия истинного автора среди кандидатов. Независимо от вида анализа, методика является устойчивой к атакам на метод и позволяет эффективно идентифицировать автора текста. Методика анализа искусственно-языкового текста [17] позволяет идентифицировать автора исходных кодов программ, учитывая специфику, отличающую их от естественно-языковых текстов, и позволяет осуществлять эффективную идентификацию вне зависимости от языка программирования, квалификации программиста и осложняющих факторов.

Модель создания автором текста в киберсреде с учетом семантики

Модель создания текста автором в киберсреде с учетом семантики представим тройкой:

$$M = (\mathbf{A},\mathbf{T},\mathbf{E}), \quad (1)$$

где $\mathbf{A}$ – множество авторов, $\mathbf{T}$ – множество текстов, $\mathbf{E}$ – множество сред.

Пусть имеется коллекция текстов $\mathbf{T} = \{t_1,...,t_{|T|}\}$ и множество авторов $\mathbf{A} = \{a_1,...,a_{|A|}\}$. Введем бинарное отношение «текст написан автором» $R \subset \mathbf{T} \times \mathbf{A}$ на декартовом произведении множеств $\mathbf{T}$ и $\mathbf{A}$ такое, что выполняется tRa если текст $t \in \mathbf{T}$ автором $a \in \mathbf{A}$:

$$\exists t \in \mathbf{T}, \exists a \in \mathbf{A} : (tRa). \quad (2)$$

В случае, когда текст t можно представить как объединение фрагментов $t = \bigcup_{i=1}^{N_t} t'_i$, написанных несколькими авторами, будем говорить, что текст t «написан в соавторстве»:

$$\exists t'_i, t'_j \subseteq t, \exists a_l, a_m \in \mathbf{A}, a_l \neq a_m : (t'_i Ra_l) \wedge (t'_j Ra_m). \quad (3)$$

Случай, когда текст, написанный одним автором, подвергается изменению другим автором при сохранении общей семантики и тональности текста назовем «редактированием» и опишем в виде функции:

$$t^e = edit(t). \quad (4)$$

Текст, полученный в результате работы генеративного алгоритма, обученного на текстах $\mathbf{T}^a$ автора a, упрощенно опишем в виде:

$$t^g = gen(\mathbf{T}^a). \quad (5)$$

Общий случай вмешательства в процесс создания текста обозначим как $inter \in \mathbf{INTER}$, где элементами множества являются факты самостоятельного написания, редактирования, соавторства, применения генеративного алгоритма.

Текст имеет семантическое описание $topic \in \mathbf{TOPIC}$. Например, такими темами порталов Интернет могут быть «новости», «язык программирования C++», «научные статьи по защите информации» и др. В свою очередь каждую тему можно представить списком ключевых слов (облаком тэгов)

$$topic_i \in \mathbf{TOPIC} = \{keyword_1,...,keyword_{|topic_i|}\}.$$

В контексте решения задач информационной безопасности будем использовать абстрактное понятие «тип текста», которое учитывает вид, стиль, жанр, назначение текста и др. $type \in \mathbf{TYPE}$. Например, в качестве типов в зависимости от задачи будем понимать художественные, любительские и сообщения из социальных сетей, естественно-языковые и искусственно-языковые и т.д.

Текст может быть отнесен к экстремистским материалам. Его действующий статус можно представить как $status \in \mathbf{STATUS}$. Возможные значения признака: разрешен или запрещен.

Каждый текст имеет эмоциональный окрас (тональность) $emo \in \mathbf{EMO}$. Эмоциональный окрас в простом случае может принимать значения: положительный, негативный, нейтральный. Возможна более детальная классификация, учитывающая оттенки радости, злости, грусти, страха, интереса и т.д.

Таким образом множество классов, к которым можно отнести текст можно описать как декартово произведение вышеозначенных множеств, а конкретный класс, к которому относится текст представить как:

$$C^t = (type,status,topic,emo) \in \mathbf{TYPE} \times \mathbf{STATUS} \times \mathbf{TOPIC} \times \mathbf{EMO} \times \mathbf{INTER}. \quad (6)$$

Каждый автор в момент создания текста представляется набором атрибутов, которые можно описать как:

$$\mathbf{C}^a = (id,age,gender,emo,status,action,pop) \in \mathbf{ID}\times\mathbf{AGE}\times\mathbf{GENDER}\times\mathbf{EMO}\times\mathbf{STATUS}\times\mathbf{ACTION}\times\mathbf{POP}, \quad (7)$$

где:

$id \in \mathbf{ID}$ – идентификатор: любая последовательность символов, которой человек себя идентифицирует (ФИО, псевдоним в социальных сетях и др.);

$gender \in \mathbf{GENDER}$ – пол и гендерная идентичность (мужской, женский, представитель ЛГБТ);

$emo \in \mathbf{EMO}$ – настроение, с которым автор писал текст (положительное, отрицательное, нейтральное);

$status \in \mathbf{STATUS}$ – статус, показывающий имеет ли человек статус иностранного агента или экстремиста, или имеет отношение к организациям, имеющим эти статусы;

$action \in \mathbf{ACTION}$ – деятельность автора по созданию текстовой информации (общение, творческая и профессиональная деятельность);

$pop \in \mathbf{POP}$ – категории популярности автора;

$age \in \mathbf{AGE}$ – возрастная группа автора.

Каждый элемент текста описывается вектором признаков, отражающим его свойства. Для естественных текстов у слова, например, можно определить часть речи, морфологические признаки и длину и т.д. Для исходных кодов программ можно определить тип токена, семантику оператора и др. Набор

признаков текста можно представить как результат работы функции извлечения полного вектора характеристик из текста:

$$\mathbf{F} = extract(t) = [f_1,...,f_{|\mathbf{F}|}], \quad (8)$$

где:

- ✓ $f_i \in \mathbf{LEX} \cup \mathbf{MORPH} \cup \mathbf{SYNT} \cup \mathbf{SEM} \cup \mathbf{IDIO} \cup \mathbf{META} \cup \mathbf{EMB}$, LEX – лексические признаки;
- ✓ MORPH – морфологические признаки;
- ✓ SYNT – синтаксические признаки;
- ✓ SEM – семантические признаки;
- ✓ IDIO – идиосинкразические признаки;
- ✓ META – метаданные текста;
- ✓ EMB – некоторое векторное представление текста, полученное с помощью модели машинного обучения, где l – размер входного слоя модели.

Множество информативных характеристик текста $\mathbf{F}^{inf}$ и их значения зависят от типа текста и от атрибутов автора, рассматриваемых в рамках интересующей задачи идентификации. Получение информативных признаков можно представить в виде функции, принимающей на вход текстовые признаки, тип текста и атрибуты автора:

$$\mathbf{F}^{inf} = inf(\mathbf{F},type,\mathbf{H}) = [f_1^{inf},...,f_{\mathbf{F}^{inf}}^{inf}], \quad (9)$$

Текстами и совокупностью векторов признаков текстов, написанных авторами, все или некоторые атрибуты которых совпадают, можно представить стиль определенной категории авторов:

$$\forall a_l,a_m \in \mathbf{A},\ \mathbf{K}^{a_l=}\mathbf{K}^{a_m} = \mathbf{K},\ \mathbf{K} \subseteq \mathbf{C}^{\mathbf{A}}:\mathbf{C}^{\mathbf{K}} = \{\mathbf{F}^{inf_i^{\mathbf{K}}}\}_{i=1}^{N_{\mathbf{T}^{\mathbf{K}}}} = \begin{cases} f^{inf_{1,1}^{\mathbf{K}}},...,f^{inf_{1,|\mathbf{F}^{inf}\mathbf{K}|}^{\mathbf{K}}}) \\ f^{inf_{N_{\mathbf{T}^{\mathbf{K}},1}}^{\mathbf{K}}},...,f^{inf_{N_{\mathbf{T}^{\mathbf{K}},|\mathbf{F}^{inf}\mathbf{K}|}}^{\mathbf{K}}}, \end{cases} \quad (10)$$

где:

- ✓ $\mathbf{K}$ – подмножество совпадающих атрибутов авторов;
- ✓ $\mathbf{T}^{\mathbf{K}}$ – множество текстов, написанных авторами с одинаковыми атрибутами;
- ✓ $N_{\mathbf{T}}^{\mathbf{K}}$ – количество текстов в этом множестве;
- ✓ $\mathbf{F}^{inf_i^{\mathbf{K}}}$ – множество значений лексических, морфологических, синтаксических, семантических, идиосинкразических признаков и эмбеддингов информативных для определенной категории авторов, имеющих одинаковые атрибуты.

Стиль автора может меняться со временем, и значения характеристик $\mathbf{F}$ могут изменяться. Однако множество информативных признаков $\mathbf{F}^{inf}$ должно быть устойчивым к этим изменениям во времени и учитывать небольшое редактирование другими авторами.

Среду, в которой автор пишет или публикует свой текст опишем как:

$$\mathbf{E} = (topic,type,status,rules,pop) \in \mathbf{TOPIC} \times \mathbf{TYPES} \times \mathbf{STATUS} \times \mathbf{RULES} \times \mathbf{POP}, \quad (11)$$

где:

- ✓ $topic \in \mathbf{TOPIC}$ – семантическое описание среды, т.е. список тематик, тексты, относящиеся к которым, размещаются в среде;
- ✓ $type \in \mathbf{TYPES}$ – тип среды. Например, Интернет-библиотека, мессенджер, социальная сеть, хостинг IT-проектов, Интернет-СМИ, сайт научного журнала, локальный компьютер автора и др.;
- ✓ $status \in \mathbf{STATUS}$ – действующий статус ресурса (разрешен или запрещен). $rules \in \mathbf{RULES}$ – правила размещения текста в среде и/или стандарты среды и необходимо ли им следовать. Например, к ним можно отнести соблюдение законодательства РФ, правил публикации статей в научном журнале, стандартов кодирования для исходных текстов программ на хостинге IT-проектов в компании, запрет публикации текстов, не соответствующих определенной возрастной категории и др.;
- ✓ $pop \in \mathbf{POP}$ – популярность среды, каждый тип среды имеет свою специфику оценки. Например, для Интернет-ресурса – количество посетителей, посетивших его за определенный период времени; для сообщества в социальной сети или канала в мессенджере – количество реальных подписчиков; сайт научного журнала – импакт-фактор и т.д.

Отметим, что чем популярнее среда, тем больший охват аудитории она имеет и тем быстрее опубликованная текстовая информация находит читателя.

Введем бинарное отношение «текст создается в среде» $Q \subset \mathbf{T} \times \mathbf{E}$ на декартовом произведении множеств $\mathbf{T}$ и $\mathbf{E}$ такое, что выполняется tQe, если текст $t \in \mathbf{T}$ создается и размещается в среде $e \in \mathbf{E}$:

$$\exists t \in \mathbf{T}, \exists e \in \mathbf{E}: (tQe). \quad (12)$$

Решение задач информационной безопасности (рис. 2) сводится к отнесению текста $t_{\mathbf{K}}$ к классу $c \in \mathbf{C} = \mathbf{C}^t \cup \mathbf{C}^{\mathbf{K}}$ с учетом ограничений и особенностей среды $\mathbf{E}$ на основе текстов, класс для которых известен $\mathbf{T}' = \{t_1,...,t_m\} \subseteq \mathbf{T}$, т.е. существует множество пар «текст-класс» $\mathbf{D} = \{t_i,c_j\}_{i=1}^m$. Целью является построение классификатора, решающего данную задачу, т.е. нахождение некоторой целевой функции $\Phi:\mathbf{T}\times\mathbf{C} \to [0,1]$. Значения функции интерпретируется как степень принадлежности объекта классу: 1 соответствует полностью положительному решению, 0 – отрицатель-

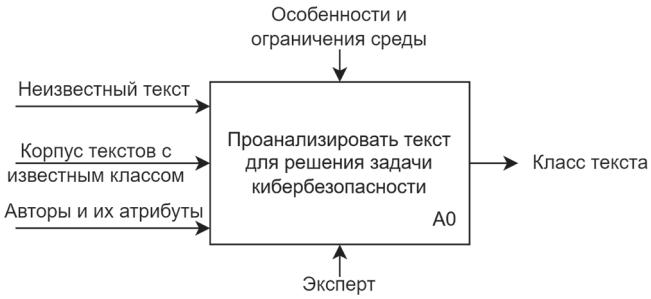

Рис. 2. IDEF0 диаграмма процесса анализа текста для решения задач кибербезопасности

Таблица 1

Примеры ограничений и влияния атрибутов автора, текста и среды

Атрибут текста	Атрибут автора	Атрибут среды	Ограничения и влияние атрибутов
t.topic	a.age	e.rules	Возрастные ограничения среды ограничивают тематику текстов, которые может публиковать или читать автор определенного возраста.
t.type	a.gender	e.topic	Некоторые среды более популярны среди определенного пола, что может влиять на выбор жанра или типа текста автором.
t.emo	a.action	e.type	Профессиональная деятельность автора коррелирует с выбором профессиональных сред и специализированных тем, что делает тексты более строгими по тональности.
t.status	a.status	e.status	Статус автора как иноагента и статус среды как разрешенного или запрещенного ресурса напрямую влияют на легальность и доступность текста для чтения.
t.status	a.action	e.rules	Авторы, занимающиеся творческой деятельностью, могут сталкиваться с ограничениями в средах со строгими правилами относительно контента.
t.topic	a.topic	e.pop	Популярные авторы выбирают тематики текстов, соответствующие интересам большинства пользователей популярной среды, чтобы увеличить свое влияние и расширить аудиторию.
t.type	a.action	e.type	Авторы-ученые часто публикуют научные работы в специализированных журналах, где тип среды соответствует типу текста.
t.status	a.status	e.rules	Авторы текстов, признанных иноагентами, могут сталкиваться с дополнительными ограничениями в определенных средах из-за строгих правил.

ному. При этом каждый текст рассматривается как вектор признаков $\mathbf{F}$.

Примеры ограничений среды, накладывающихся на авторов и создаваемые тексты, а также влияние атрибутов друг на друга приведены в таблице 1.

Интерпретация в виде графа представлена на рис. 3.

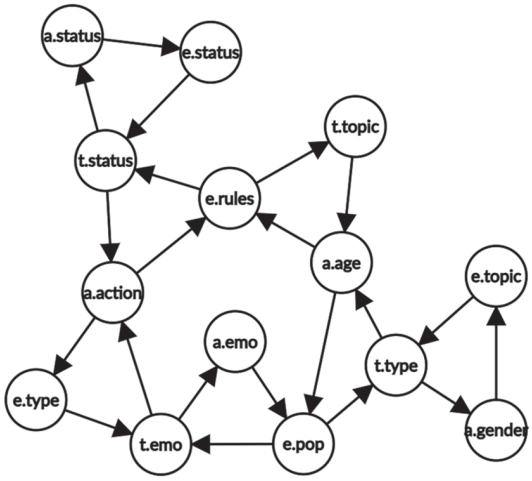

Рис. 3. Ограничения и влияние атрибутов автора, текста и среды

Заключение

В статье предлагается комплексное решение важной научной проблемы идентификации автора текстовой информации и таких авторских атрибутов, как пол и гендер, возраст, идеология и взгляды, основанное на передовых технологиях искусственного интеллекта и машинного обучения.

1. На основе анализа материалов предложена методология идентификации автора естественно-языкового текста и исходных текстов программ для решения задач информационной безопасности в виде комплекса методов, моделей и алгоритмов, агрегирующий имеющийся опыт. Методология является универсальной для решения задач информационной безопасности, связанных с классификацией текстов.

2. Предложена модель создания автором текста в киберсреде, учитывающая взаимодействие компонентов и ограничения, накладываемые на процесс личностью автора и видом деятельности, особенностями среды, текста и семантики.

3. Методология предполагает использование методов статистического анализа, машинного и глубокого обучения. Исходя из специфики текстов, решаемой задачи информационной безопасности и потенциальных атак на методы, в элементах методологии могут быть задействованы разные по принципу действия подходы. Традиционные методы машинного обучения обеспечивают высокую степень интерпретируемости и скорости, поэтому могут применяться как самостоятельно, так и в составе ансамблей методов принятия решений, однако являются менее эффективными в сравнении с НС при наличии осложняющих факторов. НС являются более устойчивыми к атакам на метод и более эффективными для поиска явных и неявных признаков авторского стиля.

4. Ключевыми в методологии являются методики идентификации автора естественно-язычного и искусственно-язычного текста, так как подразумевают

применение разных по своему принципу действия методов и инструментов. Это связано со спецификой анализируемых данных. Подходы, используемые для естественно-языкового анализа, должны обеспечивать интерпретацию неоднозначности, понимание контекста, а также разрешение двусмысленности, которые являются менее характерными для искусственно-языковых текстов. Подходы, используемые для искусственного языка, напротив, должны быть адаптированы под анализ регулярных структур, информативные признаки в которых могут быть менее выраженными ввиду следования авторами строгим синтаксическим и семантическим инструкциям и правилам.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ в рамках базовой части государственного задания ТУСУРа на 2023–2025 гг. (проект № FEWM-2023-0015).

Литература

1. *Uslu U., Durmaz Ö., Alptekin G. I. Evaluation of Deep Learning Models for Continuous Authentication Using Behavioral Biometrics // Proceedings of 27th International Conference on Knowledge Based and Intelligent Information and Engineering Sytems (KES 2023), Procedia Computer Science. – 2023. – Vol. 225. – P. 1272–1281.*
2. *Bano H, Akbar W., Aslam N., Bilal M. Identification and Classification of Extremist by Topic Modeling Sentiment Analysis // VFAST Transactions on Software Engineering. – 2023. – Vol. 11. – P. 235–248.*
3. *Аванесян Н. Л., Соловьев Ф. Н., Тихомирова Е. А., Чеповский А. М. Выявление значимых признаков противоправных текстов // Вопросы кибербезопасности. – 2020. – № 4(38). – С. 76–84.*
4. *Васильев В. И., Вульфин А. М., Кучкарова Н. В. Тематическое моделирование и суммаризация текстов в области кибербезопасности // Вопросы кибербезопасности. – 2023. – № 2(54). – С. 1–22*
5. *Araque O., Iglesias C. A. An Approach for Radicalization Detection Based on Emotion Signals and Semantic Similarity // IEEE Access. – 2020. – Vol. 8. – P. 17877–17891.*
6. *Asad M., Shafiq Z., Srinivasan P. A Girl Has A Name: Detecting Authorship Obfuscation // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics. – 2020. – P. 2235–2245.*
7. *Kovalev A. K., Kuznetsova Yu. M. Possibilities of automatic text analysis in the task of determining the psychological characteristics of the author // Experimental Psychology (Russia). – 2020. – Vol. 13, no. 1. – P. 149–158.*
8. *Москин Н. Д. Теоретико-графовые модели, методы и программные средства интеллектуального анализа текстовой информации на примере фольклорных и литературных произведений: дис. д-р. техн. наук: 05.13.18. – Петрозаводский. гос. университет, Петрозаводск, 2022. – 346 с.*
9. *Рогов А. А. Проблема атрибуции в журналах «Время», «Эпоха» и еженедельнике «Гражданин» / А. А. Рогов, Р. В. Абрамов, Д. Д. Бучнева, О. В. Захарова, К. А. Кулаков, А. А. Лебедев и др. // Издательство «Острова». – 2021. – 391 с.*
10. *Огорелков И. В. Исследование лингвистических характеристик текста с целью определения пола автора на примере анализа письменных русскоязычных текстов политического дискурса: дис. канд. техн. наук: 10.02.01. – ФГБОУ ВО «Государственный институт русского языка им. А.С. Пушкина», Москва, 2021. – 457 с.*
11. *Огорелков И. В. Исследование лингвистических характеристик письменного текста политического дискурса с целью определения пола автора // Язык. Право. Общество: сб. ст. V Междунар. науч.-практ. конф. (г. Пенза, 22–25 мая 2018 г.) / под общ. ред. О. В. Барабаш; редколегия: М. Б. Ворошилова, Т. В. Дубровская, А. К. Дятлова, Н. А. Павлова. – Пенза: Изд-во ПГУ, 2018. – 484 с. ISBN 978-5-907018-83-9. – 2018. – С. 88–93.*
12. *Сбоев А. Г. Нейросетевое моделирование и машинное обучение на основе экспериментальных и наблюдательных данных: дис. д-р. техн. наук: 05.13.18. – Национальный исследовательский центр «Курчатовский институт», Москва, 2021. – 389 с.*
13. *Sboev A. Neural Network Model to Include Textual Dependency Tree Structure in Gender Classification of Russian Text Author / A. Sboev, A. Selivanov, R. Rybka, I. Moloshnikov, D. Bogachev // Advanced Technologies in Robotics and Intelligent Systems. – Springer, Cham, 2020. – P. 405–412.*
14. *Давыдова Ю. В. Методы текстового поиска и обработки информации в социальных сетях при управлении деятельностью правоохранительных органов: дис. канд. техн. наук, 05.13.10. – ФГБОУ ВО «Орловский государственный университет имени И. С. Тургенева», Белгород, 2021. – 146 с.*
15. *Андреев И. А. Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя: дис. канд. техн. наук, 05.13.01. – Ульяновский государственный технический университет, Ульяновск, 2022. – 166 с.*
16. *Куртукова А. В., Романов А. С., Федотова А. М., Шелупанов А. А. Применение методов машинного обучения и отбора признаков на основе генетического алгоритма в решении задачи определения автора русскоязычного текста для кибербезопасности / А. В. Куртукова [и др.] // Доклады ТУСУР. – 2022. – Т. 25, № 1. – С. 79–85.*
17. *Романов А. С., Куртукова А. В., Шелупанов А. А., Федотова А. М. Идентификация автора исходного кода программы на основе неоднородных данных для решения задач кибербезопасности / А. В. Куртукова, А. А. Шелупанов, А. М. Федотова // Моделирование, оптимизация и информационные технологии. – 2022. – №10(3) [Электронный ресурс]. – URL: https://moitvivt.ru/ru/journal/pdf?id=1227 DOI: 10.26102/2310-6018/2022.38.3.016.*