Отзыв

официального оппонента на диссертационную работу
Романова Александра Сергеевича
«Методология идентификации автора текстовой информации для решения задач кибербезопасности», представленную на соискание ученой степени доктора технических наук по специальности 2.3.6 – Методы и системы защиты информации, информационная безопасность

Актуальность темы диссертации. В условиях стремительного роста объемов текстовой информации в открытых и закрытых цифровых средах и обострения киберугроз, обеспечение достоверной атрибуции текстовых материалов приобретает особую актуальность для решения задач информационной безопасности. Системы автоматической идентификации автора и сопутствующих ему признаков (пол, возраст и др.) позволяют, во-первых, оперативно распознавать источники деструктивного контента, во-вторых, в образовательной и научной сферах контролировать оригинальность и соответствие работ заявленным авторам, в-третьих, совершенствовать лингвистические методы анализа стилистических паттернов и эволюции языка.

Тематике определения авторства посвящено множество исследований, в рамках которых разработаны различные методы и подходы на базе алгоритмов статистики и машинного обучения. Однако большинство из них разработаны для английского языка и нацелены на решение прикладных задач, например, определения конкретного атрибута автора (пола или возраста) или определения авторства в рамках одного домена (социальные сети или художественная литература).

Разработка методологического и программного комплекса, интегрирующего современные алгоритмы машинного обучения для многопараметрической идентификации автора представляется важной и актуальной задачей для кибербезопасности. Она также согласуется с приоритетами Национальной стратегии научно-технологического развития РФ.

Анализ содержания диссертации. Диссертация состоит из 400 страниц печатного текста. Она включает введение, шесть глав, заключение, список литературы из 334 источников и четыре приложения.

Во введении приводятся обоснование актуальности проблематики автоматизированной атрибуции текстов для решения задач кибербезопасности, а также четкие формулировки цели и задач работы. Здесь же приводятся результаты и основные защищаемые положения, а также дается оценка научной новизны и перспектив практического применения разработанных подходов.

Первая глава посвящена обзору задач и существующих техник определения авторства как естественно-языкового текста, так и исходного кода программ. В рамках обзора выделяется ряд задач, требующих создания отдельных подходов: определение авторства, определение авторских атрибутов, продленная аутентификация пользователей социальных сетей, определение автора программного кода, выявление деструктивной направленности текста, а также различные атаки на методы. Подходы к решению этих задач сравниваются между собой, выявляются их сильные и слабые стороны. Исходя из выполненного анализа, формулируется цель работы – создание методологии идентификации автора текстовой информации, включая естественно-языковые тексты и исходные коды программ, для решения задач кибербезопасности.

Во второй главе приводится описание предложенной методологии с детальной математической проработкой модели создания автором текста в киберсреде, модели представления текстовой информации, алгоритмов текстового анализа, классификаторов, а также методов оптимизации, снижения размерности, сглаживания и отбора признаков.

Третья и четвертая главы посвящены прикладным методикам. Третья – для естественно-языковых текстов с учетом открытой и закрытой атрибуции и верификации, а четвертая – для искусственно-языковых текстов в простых и усложненных сценариях: обфускация, командная разработка, следование стандартам, генерация кода, смешанные данные. Для каждой методики приводятся результаты апробации для ряда случаев. Лучшие результаты для естественно-

языковых текстов получены гибридной моделью GRU+CNN, для верификации – One-class SVM. Лучшие результаты для исходных кодов программ получены CodeBERT с классификационным слоем. Соответствующие модели легли в основу предложенных методик определения авторства естественно- и искусственно-языковых текстов.

Пятая глава иллюстрирует применение разработанной методологии к решению важных народно-хозяйственных задач. На базе методологии формулируются прикладные методики определения деструктивной и экстремистской направленности текста, определения возраста автора (включая тексты, созданные лицами, младше 18 лет), определения пола и гендера автора, а также проверки однородности текста и поиска заимствований.

Шестая глава описывает архитектуру программного комплекса, реализующего все ключевые алгоритмы, программных продуктов, решающих частные задачи, и содержит описание подготовленных наборов данных. Отдельно хочется подчеркнуть общий объем сформированной базы данных, содержащей тексты различных доменов. Общее число текстов/кодов в наборе превышает 3,5 млн., а авторов – 1,1 млн.

В заключении резюмируются основные результаты, подтверждается выполнение всех задач и достижение цели исследования. Предложенные методики и программные решения готовы к внедрению в системы информационной безопасности и могут стать основой для дальнейших исследований в области цифровой лингвистики и форензики.

Научная новизна диссертации. Научной новизной обладают следующие результаты диссертации:

1. Предложена комплексная методология идентификации авторства, эффективная как для естественно-языковых текстов, так и для исходных кодов программ, и одновременно устойчивая к разнообразным атакам.

2. Разработана оригинальная модель создания текста в киберсреде, в которую впервые интегрированы семантические характеристики, информативные признаки

на разных уровнях анализа, специфика цифрового окружения, а также атрибуты автора и особенности его деятельности.

3. В области определения авторства естественно-языковых текстов предложена методика, сочетающая архитектуры GRU и CNN с SVM, где отбор признаков осуществляется при помощи генетического алгоритма, а надежность модели обеспечивается методами регуляризации. Эта методика впервые охватывает как открытую, так и закрытую атрибуцию, включая тексты, сгенерированные нейросетями. Результаты методики для закрытой атрибуции достигают 98,6%, 98,9%, 96,3% для художественных, любительских и коротких текстов, соответственно; для открытой – 94,2%, 98,2% и 93,5%; для верификации – 99,2%, 98,9% и 97,8%.

4. Для исходных кодов программ создана новая методика на основе глубокой модели CodeBERT, способной учитывать сложные сценарии идентификации, включая обфускацию, стандарты кодирования, искусственную-генерацию и коллективную разработку. Результаты методики достигают 93% для простого случая, а также 82% – для случая обфускации, 92% – для смешанных наборов данных, 94% – для коллективной разработки и 94% – для определения генеративной модели, создавшей код.

5. Разработана прикладная методика, направленная на классификацию авторов по возрастным группам, где для фильтрации недостоверных примеров используется компьютерное зрение, а в качестве метода принятия решения – fastText. Точность методики достигает 82%. Методика применима для разграничения возрастных групп до 18 лет и старше.

6. Разработана прикладная методика, направленная на определение экстремистского или деструктивного характера текста и идентификацию его автора на базе семантической кластеризации, гибридных моделей GRU+CNN и BERT с применением трансферного обучения. Точность методики достигает 95%.

7. Разработана прикладная методика, направленная на определение пола и гендера автора русскоязычного текста с учетом ЛГБТ-гендеров. Методика основывается на ансамбле SVM, обученной на признаках, полученных

семантической кластеризацией, CNN, обученной на сглаженных методом Катца распределениях триграмм, и BERT. Точность методики достигает 93%.

8. Разработана прикладная методика, предназначенная для анализа однородности текста и поиска заимствований. Она основана на сиамских сетях с тройной и контрастивной функциями потерь, и впервые позволяет решать задачи обнаружения плагиата в условиях открытого множества возможных авторов и при наличии искусственно сгенерированных фрагментов. Точность методики достигает 94% для оценки однородности и 99% – для подтверждения авторства.

Теоретическая значимость полученных результатов. Представленные в диссертации результаты вносят существенный вклад в теорию идентификации автора текстовой информации, поскольку в исследовании проведена глубокая систематизация классических лингвистических и компьютерно-лингвистических исследований и на этой основе выстроена единая методология анализа авторского стиля. Диссертация предлагает новые принципы обработки и анализа текстов, позволяющие не только надежно идентифицировать автора, но и оценивать его атрибуты, обнаруживать аномалии и потенциальные риски. Комплексное и формально обоснованное сочетание лингвистических знаний и передовых технологий искусственно интеллекта делает исследование теоретически значимым и перспективным для решения множества прикладных задач в области информационной безопасности.

Практическая значимость полученных результатов. Разработанные методики могут быть использованы в областях цифровой криминалистики, правоохранительной деятельности, защиты интеллектуальной собственности и авторских прав и др. В частности, предложенные методики могут быть использованы для решения таких задач как:

– атрибуция авторства экстремистских и деструктивных текстов;

– выявление информационных угроз, связанных с манипуляцией общественным мнением, в социальных сетях;

– сбор и формирование доказательных баз для расследований инцидентов ИБ;

– определение авторства вредоносного ПО;

– защита интеллектуальной собственности и авторского права;

– проверка оригинальности научных и учебных работ;

– проверка на факт использования ИИ-генерированного контента.

Обоснованность и достоверность научных положений, выводов и рекомендаций. В диссертации последовательно рассмотрены все ключевые этапы исследования: постановка задач, разработка методик и моделей, их теоретическое обоснование, программная реализация и апробация на практических примерах. Надежность и воспроизводимость полученных результатов подтверждаются:

– корректным выбором и обоснованием статистических, эвристических и нейросетевых методов обработки данных;

– экспериментальной оценкой на больших корпусах текстов и исходного кода;

– сопоставлением с результатами других научных коллективов;

– успешным внедрением решений в ФГАОУ ВО «ТУСУР», ООО «НТР Томск», ООО «СИБ», ООО «Сибэдж», ООО «НИЦ», ФГАОУ ВО НИ ТПУ, УМВД России по Томской области, войсковую часть 51952 и на экономический факультет МГУ им. М. В. Ломоносова (подтверждено актами и отзывами).

Основные итоги исследования отражены в 31 печатной работе: 18 статей в журналах ВАК, 13 публикаций в изданиях, индексируемых Web of Science и Scopus; получено 8 свидетельств о регистрации программ для ЭВМ и 3 – о регистрации баз данных.

Замечания и вопросы по диссертации

1. При описании разработанной методологии (в автореферате на рисунке 3) методы снижения размерности включают в себя методы отбора признаков тоже. Почему предложена подобная группировка?

2. В п. 3.3.4 диссертации при использовании генеративных моделей для создания текстов на основе авторских, большую роль играет формат и структура системного и пользовательского запроса к большой языковой модели. Вопрос создания подобного запроса и его влияния на результат не раскрыт.

3. В п. 4.5 диссертации на рисунке 4.4 приведен блок «Обучение генеративных моделей GPT-2, 3, 4 на полном датасете исходных кодов», но не раскрыты подробности данного процесса с применением больших языковых моделей.

4. В п. 5.2.2.2 при описании алгоритма фильтрации недостоверных данных о возрасте с помощью предобученной нейросетевой модели VGG-Face не раскрыты подробности обработки deepfake или обработанных с помощью специализированных моделей фотографий.

5. При анализе угроз информационной безопасности в контексте идентификации авторов противоправных текстов уместной была бы привязка к действующей Методике оценки угроз безопасности информации ФСТЭК России и к БДУ ФСТЭК, в частности, УБИ.10 «Угроза распространения противоправной информации», с целью конкретизации способов применения разработанных решений для противодействия реализации данной угрозы.

Отмеченные недостатки не затрагивают сущности научных положений, выносимых на защиту, и не влияют на общую оценку работы, которая, несомненно, положительна.

Заключение

Диссертационная работа Романова А.С. представляет собой законченную научно-квалификационную работу, написана на актуальную тему, выполнена на высоком научном уровне, отличается научной новизной и практической значимостью, выполнена на высоком научно-техническом уровне. Автором в диссертации сформулирована и решена научная проблема, имеющая важное хозяйственное значение в области кибербезопасности, компьютерной лингвистики и криминалистически для идентификации автора естественно-языкового текста и текстов программ. Решение данной проблемы имеет научную и практическую ценность для построения эффективных систем защиты информации. Внедрение разработанных методологии, методик и программных систем вносит значительный вклад в развитие кибербезопасности страны.

Результаты работы достаточно полно представлены в публикациях автора. Автореферат полностью соответствует содержанию диссертации.

Тема диссертационной работы соответствует паспорту специальности 2.3.6 – Методы и системы защиты информации, информационная безопасность: пункты 1, 5, 12, 13.

Диссертация Романова Александра Сергеевича удовлетворяет всем требованиям пп. 9-14 «Положения о порядке присуждения ученых степеней» ВАК РФ, утвержденного постановлением Правительства РФ от 24.09.13 №842 (редакция от 16.10.2024), предъявляемых к докторским диссертациям, а ее автор заслуживает присуждения ученой степени доктора технических наук по специальности 2.3.6 – Методы и системы защиты информации, информационная безопасность.

Я, Вульфин Алексей Михайлович, даю свое согласие на включение своих персональных данных в документы, связанные с работой диссертационного совета, и их дальнейшую обработку.

Официальный оппонент,
доктор технических наук, профессор кафедры вычислительной техники и защиты информации Федерального государственного бюджетного образовательного учреждения высшего образования «Уфимский университет науки и технологий»
450076, Республика Башкортостан, г.о. город Уфа, г. Уфа, ул. Заки Валиди, д. 32
Тел. 8(347)272-63-70
e-mail: vulfin.am@ugatu.su

Алексей Михайлович Вульфин

«09» сентября 2025

Диссертация на соискание ученой степени доктора технических наук защищена в 2022 году по специальности 2.3.6 – Методы и системы защиты информации, информационная безопасность.

Подпись Вульфина А.М.
Удостоверяю «09» 09 2025 г.
Начальник общего отдела УУНиТ
Рахимова Ф.