Научная статья
УДК 004.056.53
DOI:10.31854/1813-324X-2024-10-2-92-101

Подход к обнаружению вредоносных ботов в социальной сети ВКонтакте и оценка их параметров

Андрей Алексеевич Чечулин [1, 2] ✉, chechulin.aa@sut.ru
Максим Вадимович Коломеец [3], maksim.kalameyets@newcastle.ac.uk

[1] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация
[2] Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича, Санкт-Петербург, 193232, Российская Федерация
[3] Ньюкаслский университет, Ньюкасл-апон-Тайн, NE4 5TG, Соединенное Королевство

Аннотация: *Появление новых разновидностей ботов в социальных сетях и совершенствование их возможностей имитации естественного поведения реальных пользователей представляют собой актуальную проблему в области защиты социальных сетей и онлайн сообществ. В данной работе предлагается новый подход к обнаружению и оценке параметров ботов в рамках социальной сети ВКонтакте. Основой предложенного подхода является создание наборов данных с использованием метода «контрольной покупки» ботов, который позволяет оценить такие характеристики как стоимость, качество и скорость действия ботов, а с использованием теста Тьюринга также насколько пользователи доверяют ботам. В совокупности с общепринятыми методами машинного обучения и признаками, извлеченными из графов взаимодействий, текстовых сообщений и статистических распределений, становится возможным достаточно точно не только обнаруживать ботов, но и предсказывать их характеристики. В работе демонстрируется, что итоговая модель, построенная на основе предлагаемого подхода, робастна к разбалансированным данным и может идентифицировать большинство видов ботов, так как имеет лишь незначительную корреляцию с их основными характеристиками. Предложенный подход может использоваться в рамках выбора контрмер для защиты социальных сетей и исторического анализа, позволяя не только подтвердить присутствие ботов, но и характеризовать специфику атаки.*

Ключевые слова: *безопасность социальных сетей, социальные боты, социальная инженерия, метрики, дезинформация, фейковые аккаунты, анализ рисков*

Источник финансирования: Работа Андрея Чечулина была профинансирована в рамках бюджетного проекта FFZF-2022-0007. Максим Коломеец не получал финансирования в рамках данного исследования.

Ссылка для цитирования: Чечулин А.А., Коломеец М.В. Подход к обнаружению вредоносных ботов в социальной сети Вконтакте и оценка их параметров // Труды учебных заведений связи. 2024. Т. 10. № 2. С. 92–101. DOI:10.31854/1813-324X-2024-10-2-92-101. EDN:ZTCHLS

Approach to Detecting Malicious Bots in the Vkontakte Social Network and Assessing Their Parameters

Andrey Chechulin [1, 2] ✉, chechulin.aa@sut.ru
Maxim Kolomeets [3], maksim.kalameyets@newcastle.ac.uk

[1] St. Petersburg Federal Research Center of the Russian Academy of Sciences,
St. Petersburg, 199178, Russia
[2] The Bonch-Bruevich Saint-Petersburg State University of Telecommunications,
St. Petersburg, 193232, Russian Federation
[3] Newcastle University,
Newcastle upon Tyne, NE4 5TG, UK

Abstract: *The emergence of new varieties of bots in social networks and the improvement of their capabilities to imitate the natural behavior of real users represent a significant problem in the field of protection of social networks and online communities. This paper proposes a new approach to detecting and assessing the parameters of bots within the social network «VKontakte». The basis of the proposed approach is the creation of datasets using the method of «controlled purchase» of bots, which allows one to assess bots' characteristics such as price, quality, and speed of action of bots, and using the Turing Test to assess how much users trust bots. In combination with traditional machine learning methods and features extracted from interaction graphs, text messages, and statistical distributions, it becomes possible to not only detect bots accurately but also predict their characteristics. This paper demonstrates that the trained machine learning model, based on the proposed approach, is robust to imbalanced data and can identify most types of bots as it has only a minor correlation with their main characteristics. The proposed approach can be used within the choice of countermeasures for the protection of social networks and for historical analysis, which allows not only to confirm the presence of bots but also to characterize the specifics of the attack.*

Keywords: *social media security, social bots, social engineering, metrics, disinformation, fake accounts, risk analysis*

Funding: *Andrey Chechulin's work was funded under the budget project FFZF-2022-0007. Maxim Kolomeets received no funding for this study.*

For citation: Chechulin A., Kolomeets M. Approach to Detecting Malicious Bots in the Vkontakte Social Network and Assessing Their Parameters. *Proceedings of Telecommunication Universities.* 2024;10(2):92–101. (in Russ.) DOI:10.31854/1813-324X-2024-10-2-92-101. EDN:ZTCHLS

1. Введение

Противодействие ботам в социальных сетях является актуальной областью исследований, учитывая заметное влияние разного рода вредоносных аккаунтов на различные социально значимые процессы и повседневную жизнь пользователей. Данная проблема усугубляется феноменом «эволюции ботов», когда атакующие постоянно совершенствуют подходы к созданию все более правдоподобных аккаунтов, которые могут имитировать естественное поведение пользователей [1]. Катализаторами эволюции ботов безусловно стали Генеративные Состязательные Сети (GAN, *аббр. от англ.* Generative Adversarial Network), которые позволяют ботам генерировать изображения [2] и Большие Языковые Модели (LLM, *аббр. от англ.* Large Language Models), облегчающие генерацию текстового контента [3]. В результате новые разновидности ботов обладают высокой степенью автоматизации и часто могут имитировать человеческое поведение настолько точно, что становится практически невозможно отличить их от реальных пользователей [4]. В том время как большинство исследований по обнаружению ботов посвящено Twitter [5], в данной работе представлено решение для идентификации ботов в социальной сети ВКонтакте – российской онлайн платформе, насчитывающей более 70 миллионов ежемесячно-активной аудитории. Предлагаемый подход выходит за рамки одного лишь обнаружения – он также позволяет получить оценку различных параметров ботов [6], которые предоставляют информацию о способностях атакующего, включая *стоимость* атаки, *качество* ботов, уровень *доверия* пользователей к ботам, *скорость* выполнения атаки и *тип продавца* ботами.

Новизна данной работы заключается в подходе, позволяющем оценивать параметры ботов, а также в схеме построения признаков для обнаружения ботов социальной сети ВКонтакте, включающей анализ числовых распределений, графов взаимодействий пользователей и текстового содержания. Предложенный подход позволяет идентифицировать ботов, различая не просто «бинарное присутствие», но и характеризовать атаку через анализ параметров идентифицированных аккаунтов. Подобная детальная характеристика может существенно повлиять на выбор контрмер и качественно улучшить мониторинг активности ботов во ВКонтакте.

Данная статья содержит обзор текущего состояния исследований по тематике обнаружения ботов в ВКонтакте; описывает первую (формирование наборов данных с использованием методики закупки и теста Тьюринга) и вторую часть подхода

(построение признаков и обучение модели); результаты анализа информативности признаков; анализ возможности модели (эффективность обнаружения ботов, эффективность оценки параметров ботов и то, какие параметры ботов влияют на возможность обмана детектора); обсуждение результатов экспериментов; выводы.

2. Контекст научной проблемы и релевантные исследования

Большинство существующих исследований по обнаружению ботов в социальной сети ВКонтакте основаны на методах машинного обучения с использованием данных, извлеченных из профилей пользователей графов их друзей или подписчиков.

В исследовании [7] авторы использовали нейронную сеть прямого распространения в сочетании с признаками, отражающими общую информацию об аккаунте (такую как возраст, количество фотографий и т. д.), и признаками на основе анализа черных списков, содержащих URL-адреса и определенные фразы, упомянутые в описаниях аккаунтов.

Другое исследование [8] посвящено анализу информативности категориальных признаков аккаунтов ботов с использованием классификатора Catboost.

Исследование, описанное в [9], использует в качестве признаков информацию о «полноте» профиля аккаунта – степени заполненности анализируемых полей аккаунта. «Полнота» профиля в совокупности с данными, извлеченными из списков друзей и подписчиков, служили входным вектором для детектора ботов на основе алгоритма случайного леса.

Одно из предыдущих авторских исследований [10] посвящено обнаружению ботов, основываясь исключительно на списках друзей, без какого-либо анализа самого профиля. Такой подход может быть полезен для анализа аккаунтов, ограничивающих доступ к своему профилю настройками приватности, так как список друзей можно косвенно установить путем поиска целевого аккаунт в списках друзей других пользователей. Однако такой подход требует сбора информации о списках друзей всех пользователей ВКонтакте.

Один из наиболее интересных подходов представлен в [11], где авторы предложили стековый ансамбль из нескольких классификаторов, каждый из которых использует разные признаки, такие как текст, меры центральности графов и вложения графов (embeddings). Авторы также отмечают исключительную информативность признаков графов дружбы, утверждая, что *«создание бота с графом дружбы, похожим на обычного пользователя, является сложной и время затратной задачей»*, поэтому графы дружбы имеют более высокую предсказательную способность.

Одним из недостатков вышеупомянутых исследований является их зависимость от методов маркировки, в частности – *метода на основе блокировки аккаунтов*. Данный метод основан на сборе информации о том, какие аккаунты были заблокированы социальной сетью. Это приводит к тому, что обученный классификатор не сможет превзойти эффективность встроенного обнаружения ботов ВКонтакте, а также интегрирует любые неточности из системы обнаружения ботов ВКонтакте в обучаемые классификаторы.

Другим недостатком является то, что рассмотренные подходы представляют собой двоичные результаты обнаружения (бот/не бот) и не включают компоненты качественного анализа идентифицированных ботов. В данной работе представлен подход, который дополняет общепринятые системы обнаружения в контексте вышеназванных недостатков – предлагаемый подход не только эффективно обнаруживает ботов, но и описывает характеристики аккаунтов и, следовательно, особенности атаки.

3. Предлагаемый подход – формирование наборов данных

Цель предлагаемого подхода выходит за рамки простого обнаружения ботов или порождаемой ими злонамеренной активности – подход помогает характеризовать уровень угрозы идентифицированных аккаунтов ботов с использованием *количественных метрик*. В этом разделе описывается процесс формирования набора данных, используемого для обучения модели, способной их предсказать. Применяемая в настоящей работе методология формирования наборов данных была ранее представлена в работе авторов [6]. В данном разделе кратко изложены ее основные компоненты, позволяющие получить наборы с количественными метриками ботов: а) маркировку на основе покупки ботов и б) маркировку, на основе теста Тьюринга.

Концепция маркировки на основе покупки основана на организации контролируемой атаки (злонамеренной деятельности ботов) на специально созданные для сбора аккаунты – ловушки (honeypots). Данные аккаунты-ловушки имитируют жертву, не вызывающую подозрений у продавцов ботами, что позволяет: а) заказать атаку у продавцов ботами; б) идентифицировать аккаунты, взаимодействующие с жертвой как ботов; в) измерить метрики ботов во время атаки. После атаки, с использованием теста Тьюринга определяются метрики, отражающие умение пользователей различать идентифицированных ботов (реализуя одну из двух представленных ниже методик).

Методика 1. Данный способ включает создание ловушки в виде фальшивого сообщества в социальной сети, которое выглядит реальным, но не привлекает внимание настоящих пользователей,

чтобы предотвратить их взаимодействие с ловушкой (например, служба такси в несуществующем городе). Данное сообщество используется в качестве жертвы – у продавца ботами заказываются услуги по «продвижению» либо «дискредитации» сообщества. После успешного выполнения заказа, в набор данных сохраняются идентификаторы ботов (все провзаимодействовавшие с ловушкой аккаунты), предоставленная продавцом информация (тип продавца ботами, качество бота и цена), а также скорость выполнения заказа (скорость ботов). В данной работе, в качестве покупки использовались лайки на публикации сообщества (от 100 до 300 лайков). После сбора необходимой информации деятельность ботов удаляется из сообщества, и процесс повторяется для разных продавцов и уровней качества ботов для формирования как можно более разнообразного набора. Таким образом, с помощью данной методики можно получить наборы аккаунтов, участвовавших в атаках, и для каждого такого набора определить: качество бота, тип продавца бота, цену и скорость.

Методика 2. Оценка способности человека к распознаванию ботов проводилась путем расчета метрики «Доверия» с использованием теста Тьюринга. Аннотаторам поручалось маркировать ботов, и разница между их ответами и фактической классификацией (бот/не бот) служила индикатором способности человека к обнаружению бота. В эксперименте аккаунты настоящих пользователей состояли из случайно выбранных пользователей ВКонтакте, активных пользователей различных сообществ и проверенных аккаунтов студентов. В качестве ботов использовались аккаунты, собранные на предыдущем этапе (покупка атаки). Аннотаторам представлялась коллекция из 101 аккаунта, которые они должны были категоризировать, как а) бота; б) настоящего пользователя или в) не определено. После этого подсчитывается, сколько ботов из каждого набора было правильно распознано, и формируется метрика «Доверие» как коэффициент истинно положительных результатов – соотношение правильно идентифицированных ботов к общему количеству ботов в наборе. Этот набор данных доступен в открытом доступе на GitHub [12] и включает 18 444 уникальных идентификатора ботов на основе 69 предложений (атак) 29 продавцов ботами вместе с рассчитанными метриками ботов (представлены в таблице 1).

4. Предлагаемый подход – обучение моделей

Обучение модели следует общепринятой схеме анализа данных, но с достаточно большим этапом формирования признаков. Для извлечения признаков ботов предлагается использовать сразу несколько различных источников данных об аккаунте, таких как: а) числовые и временные распределения; б) графы взаимодействия и в) текстовый контент.

ТАБЛИЦА 1. Описание метрик ботов

TABLE 1. Bot Metrics Description

Метрика	Описание метрики	Диапазон
Цена	Стоимость действия бота	$[0, \infty)$ рублей
Качество	Качество аккаунта исходя из описания продавца	{0=НИЗКОЕ, 1=СРЕДНЕЕ, 2=ВЫСОКОЕ}
Тип продавца	Разновидность совершения закупки активности ботов	{0=МАГАЗИН, 1=БИРЖА}
Скорость	Скорость действия аккаунта	{0=МИНУТА, 1=ЧАС, 2=ДЕНЬ}
Доверие	Вероятность распознавания бота пользователем	[0,1]

Числовые распределения, формирующиеся исходя из общедоступных данных об анализируемом аккаунте, можно представить в виде распределений: распределение лайков, комментариев, а также количество друзей, подписок, публикаций и лайков среди друзей и подписчиков. Для извлечения признаков из численных распределений предлагается использовать методы, описанные в таблице 2 (раздел А).

Временные распределения формируются из временных меток (например, распределение временных меток публикаций, комментариев, фотографий). Для извлечения признаков предлагаются методы, перечисленные в таблице 2 (раздел Б).

Для формирования признаков из *графов взаимодействий*, предлагается формировать списки аккаунтов, взаимодействующих с аккаунтом бота (граф, обозначенный красным на рисунке 1а).

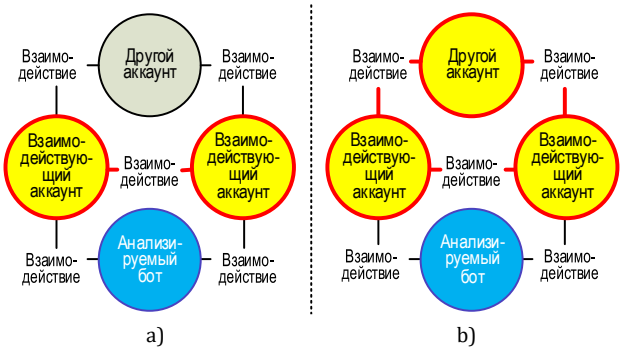


Рис. 1. Формирование двух вариантов графов взаимодействий: а) обычные; b) расширенные

Fig. 1. Construction of Two Types of Interaction Graphs: a) Conventional; b) Expanded

Данные «взаимодействующие аккаунты» служат вершинами графа и могут представлять собой следующие: а) список друзей; б) список подписчиков; в) аккаунты, которые оставляли лайки; г) оставляли комментарии.

Ребра между этими аккаунтами представляют один из следующих видов взаимодействий: а) взаимные дружеские отношения; б) подписки; в) обмен лайками; г) обмен комментариями.

ТАБЛИЦА 2. Алгоритмы извлечения признаков

TABLE 2. Feature Extraction Algorithms

Группа алгоритмов	Алгоритмы	Кол-во признаков
А – числовые распределения		
Базовые статистики (БС)	размер, min/max, среднее (стандартное, геометрическое, гармоническое), мода, отклонение, дисперсия, $q1$–$q9$	19
БС с удаленными хвостами		19
Другие	Индекс Джини, первые цифры ($q1$–$q9$) & p-значение соответствия закону Бенфорда	11
Б – Временные распределения		
P времени активности, ч	0–4 ... 20–24	6
P дня активности	ПН ... ВС	7
БС распределений	диапазоны активности	19
БС распределений	диапазоны неактивности	19
В – Графы		
Коэффициенты	% изолированных вершин, размер K-shell, размер ядра, S метрика, (N/максимальный размер/модулярность сообщества) на основе алгоритма модулярности/распространения метки (PM), % вершин доминирующего/независимого множества, коэффициент ассортативности, среднее кластерности, N мостов, коэффициент эффективности	15
Меры центральности	БС распределений: id вершин, степень связности, PageRank, VoteRank, размеры сообществ на основе модулярности/PM, размеры k-ядер	152
Г – текстовый контент		
Эмодзи	N, N графем	2
Тональность	негативный, нейтральный, неопределенный, позитивный, цитата	5
Частотность частей речи	Существительное, ..., Междометие	18
Частотность символов и слов	N латин./кирил. (символов/предложений/слов/аббревиатур/алфавитно-цифровых слов), количество пунктуаций/хэштегов/упоминаний/ссылок/емейлов/телефонов	18

Таким образом, с помощью различных комбинаций можно сгенерировать 16 различных типов графов, которые будут топологически отличаться друг от друга (4 типа вершин × 4 типа ребер).

Расширенные графы включают не только «взаимодействующие аккаунты», но и любой аккаунт, имеющий связи хотя бы с двумя другими «взаимодействующими аккаунтами» (отмечен красным на рисунке 1b). Расширенные графы предлагают альтернативный взгляд на динамику вокруг аккаунта бота, формируя графы с более сильной связью и связанной топологией.

Для формирования признаков из графов взаимодействий и расширенных графов, предлагается использовать алгоритмы, перечисленные в таблице 2 (раздел В).

Источниками данных *текстового контента* служили списки публикаций (постов), автором которых является бот, а также комментариев. Для извлечения признаков предлагается использовать алгоритмы, указанные в таблице 2 (раздел Г). Важно отметить, что для извлечения признаков не применяются методы на основе глубокого обучения (например, Трансформеры), которые могли бы интерпретировать тематическое содержание текста. Это связано с тем, что атакующие часто используют ботов в конкретных социальных контекстах, таких как атака на определенный продукт, компанию или человека. Следовательно, существует риск того, что модель может ложно классифицировать все аккаунты, обсуждающие определенную тему, как ботов. По этой причине были выбраны методы, которые основаны на статистике и синтаксическом анализе, и которые не способны учитывать смысловое содержание и, таким образом, имеют меньший риск предвзятости [13].

Методы анализа текстовых признаков используются совместно с методами анализа распределений, как показано на рисунке 2.

Учитывая, что текст в социальных сетях проявляется в виде дискретных сущностей (публикации, комментарии), вектор текстовых признаков формируется для каждой отдельной сущности. Данная процедура генерирует матрицу, где строки соответствуют сущностям, а столбцы – текстовым признакам (отмечено желтым на рисунке 2).

Для получения признаков на уровне аккаунта, а не на уровне сущности, они рассчитываются с использованием методов на основе анализа распределений для каждого столбца этой матрицы (относящегося к текстовому признаку) (отмечено серым на рисунке 2).

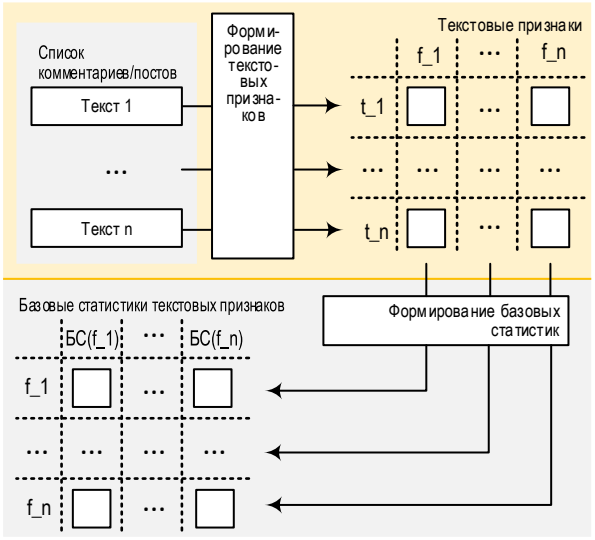


Рис. 2. Схема формирования признаков из текстовых данных

Fig. 2. Feature Construction Schema for Text Data

Полученные признаки являются вектором текстовых признаков для аккаунта, тем самым преобразуя всю матрицу в одномерный вектор. Необходимо отметить, что в данной работе также был использован один признак на основе фотографии аккаунта, указывающий на то, использует ли анализируемый аккаунт стандартное фото. Итоговая схема извлечения признаков подробно описана в таблице 2. Чтобы минимизировать сложность модели, предлагается использовать корреляцию Спирмена для выявления признаков, которые показали наиболее сильную предсказательную связь с целевыми метками (подробнее в разделе «Информативность признаков»). Полученный набор признаков впоследствии используется в качестве входного вектора нейронной сети в прямой связи для решения проблемы бинарной классификации (бот/пользователь) или регрессионного анализа для предсказания метрик ботов.

5. Информативность признаков

В результате операции извлечения было сформировано 234 467 признаков. Для упрощения модели путем уменьшения размерности пространства признаков предлагается использовать двухэтапную схему выбора:

1) устранение мультиколлинеарных признаков, когда корреляция Спирмена между признаками превышает 0,5;

2) идентификация топ-1000 признаков, показывающих наиболее сильную корреляцию с целевой меткой (бот/пользователь).

Анализ мультиколлинеарности сократил количество признаков до 21 574. Кроме того, была оценена информативность отобранных признаков на основе их корреляции с целевой меткой, где 0 обозначает настоящего пользователя, а 1 представляет бота. Рисунок 3 иллюстрирует распределение информативности признаков по источникам данных для собранных наборов ботов и 100 000 случайных пользователей ВКонтакте. Очевидно, что наиболее эффективными признаками для обнаружения ботов были те, которые были получены из графов и расширенных графов. Напротив, признаки, основанные на распределениях, были наименее информативными, что соответствует эмпирическим наблюдениям – боты часто имитируют числовые данные своих профилей в попытке ввести пользователей в заблуждение.

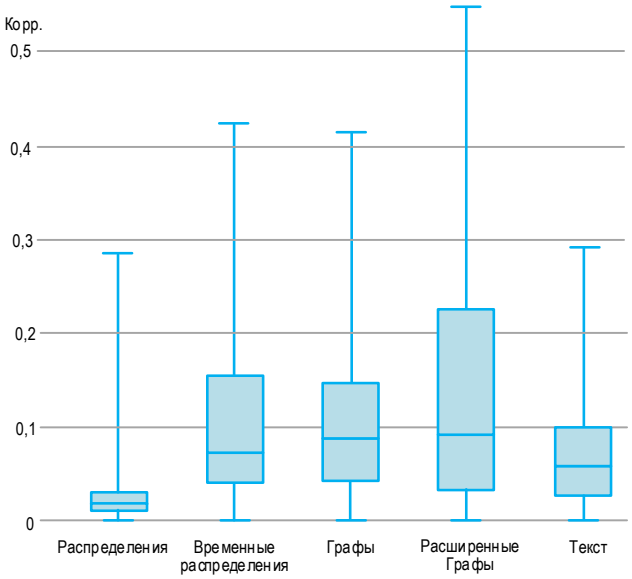


Рис. 3. Распределение информативности (ось *Y*) признаков по источникам данных (ось *X*)

Fig. 3. Informativeness Distribution (Y Axe) of Features for Data Types (X Axe)

6. Эффективность обнаружения ботов, предсказания метрик и оценка возможности обмана детектора

Для оценки эффективности классификатора в задаче идентификации отдельных ботов были использованы активные аккаунты выборки из 100 000 случайных пользователей социальной сети ВКонтакте в качестве *отрицательных* (Истинно Отрицательные: True Negative – *TN*; Ложно Положительные: False Positive – *FP*). Для *положительных* (Истинно Положительные: True Positive – *TP* и Ложно Отрицательные: False Negative – *FN*) были использованы активные аккаунты из числа 18 444 ботов из собранных наборов данных. Модель обучалась на 70 % аккаунтов и тестировалась на оставшихся 30 %.

Метрики, описывающие эффективность обнаружения ботов, перечислены в таблице 3, а график площади под ROC-кривой (AUC-ROC) изображен на рисунке 4а. Чтобы оценить, все ли наборы ботов успешно детектируются моделью, был рассчитан коэффициент истинно положительных результатов (TPR, *аббр. от англ.* True Positive Rate) для каж-

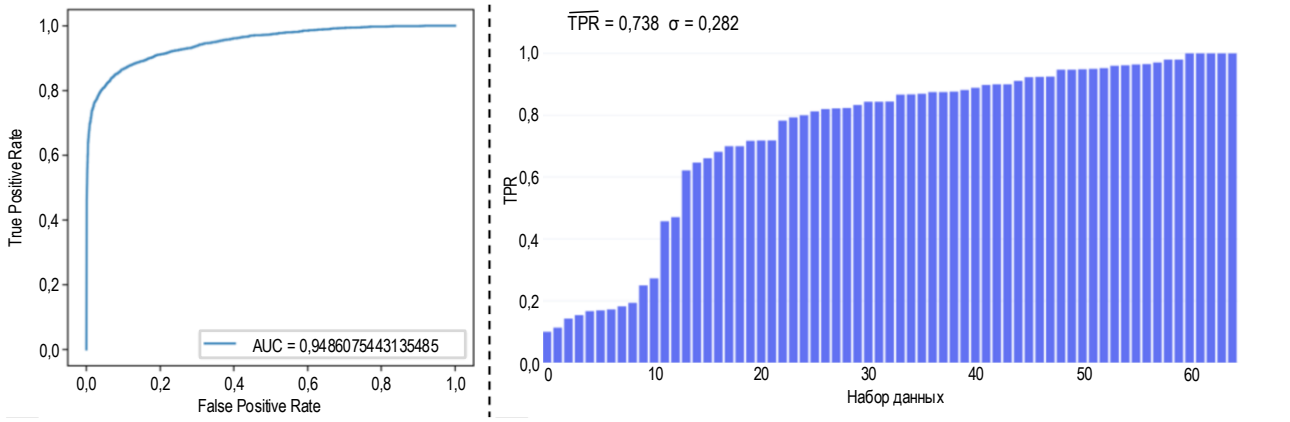
дого набора ботов. Распределение *TPR* по наборам, а также его среднее значение и стандартное отклонение, представлены на рисунке 4b.

Для метрик ботов была оценена средняя абсолютная ошибка (MAE, *аббр. от англ.* Mean Absolute Error) прогнозов обученной модели (таблица 5). На рисунке 5 представлены распределения MAE (слева) по различным наборам типов ботов, а также предсказанные и истинные значения их метрик (посредине). Чтобы определить, какие именно боты обладают большей способностью обходить обученную систему обнаружения, была исследована взаимосвязь между *TPR* и метриками ботов. Корреляция между *TPR* и метриками наборов ботов представлена на рисунке 5 (справа). Коэффициенты корреляции Спирмена, которые отражают связь между *TPR* и различными метриками, перечислены в таблице 4.

ТАБЛИЦА 3. Эффективность детектирования ботов в соответствии с рисунком 4

TABLE 3. Bot Detection Efficiency According to Figure 4

TP	*TN*	*FP*	*FN*	*Prec.*	*Rec./TPR*	*TNR*	*AUC*
2500	7966	160	819	0,940	0,753	0,980	0,94



a) b)

Рис. 4. Результаты оценки качества детектирования ботов: a) AUC-ROC, где экземпляр данных – это отдельный аккаунт; b) распределение TPR, где каждый столбец – это набор ботов определенного вида

Fig. 4. Results of Bot Detection Quality Evaluation: a) AUC-ROC Where Data Example is a Single Account; b) TPR Distribution Where Each Column is a Set with Bots of Specific Type

ТАБЛИЦА 4. Средняя абсолютная ошибка предсказания метрик/корреляция между TPR (способность обнаружить бота) и метриками (характеристики бота) в соответствии с рисунком 5

TABLE 4. Mean Absolute Error of Metrics Prediction/ Correlation between TPR (Ability to Detect Bot) and Metrics (Bot Characteristics) According to Figure 5

Метрика	Качество	Доверие	Скорость	Цена	Тип продавца
MAE/корреляция Спирмена	0,297/–0,172	0,126/0,477	0,226/–0,140	0,186/–0,021	0,136/–0,245

7. Обсуждение результатов

Результаты экспериментов показывают, что использование общепринятых методов машинного обучения в сочетании с построением признаков на основе графов, текста и распределений, а также формированием наборов данных на основе метода покупки, позволяет эффективно обнаруживать ботов и оценивать их параметры. Экспериментальная оценка обученной модели позволила сделать следующие выводы.

Во-первых, AUC-ROC модели обнаружения составляет 0,949 с *TPR* = 0,753, как показано в таблице 3 и на рисунке 4a. Модель показывает высокий *TNR* = 0,980, что говорит о том, что детектор сохраняет точность даже на несбалансированных входных данных, когда в анализируемых данных настоящих пользователей значительно больше, чем ботов.

Во-вторых, анализ среднего *TPR* по различным наборам ботов, как показано на рисунке 4b, равен 0,738. Очевидно, что эффективность детектора отличается для различных наборов ботов, при этом только 10 наборов ботов имеют *TPR* около 0,2, что свидетельствует о том, что существуют определенные разновидности ботов, которые достаточно эффективно могут избежать обнаружения. Тем не менее, для большинства наборов ботов *TPR* приближается к 0,8.

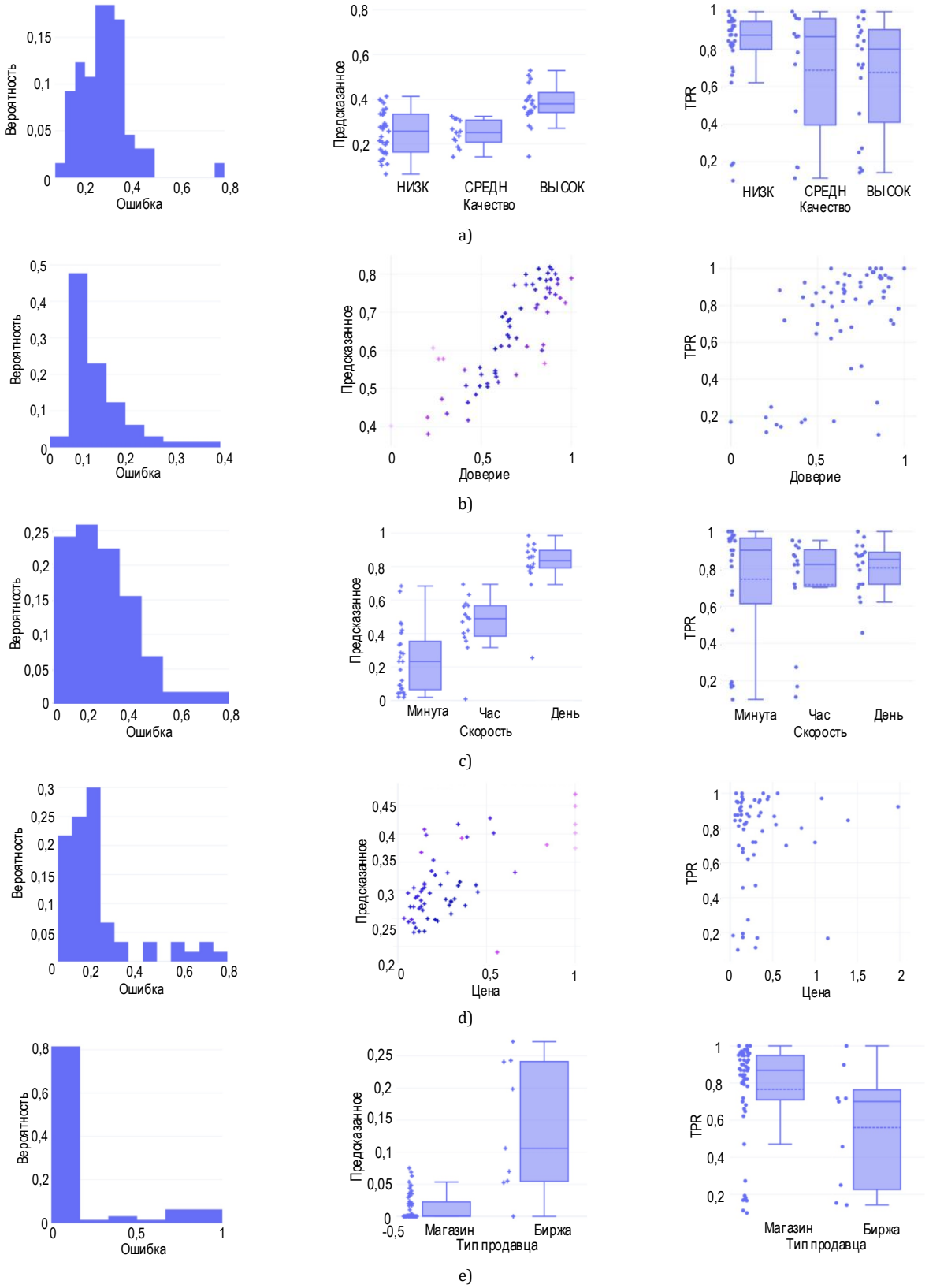


Рис. 5. Результаты оценки эффективности детектирования ботов (распределение MAE – слева, график истинных/ предсказанных значений – посредине, зависимость TPR от метрик ботов – справа) для различных метрик: a) качество; b) доверие; c) скорость; d) цена; e) тип продавца

Fig. 5. Results of Bot Detection Efficiency (MAE Distribution – Left Side, Chart Actual/Predicted Values – Middle, TPR Dependency on Bot Metrics – Right Side) for Different Metrics: a) Quality; b) Trust; c) Speed; d) Price; e) Bot Trader Type

В-третьих, согласно таблице 4, предсказание метрик ботов достаточно точно (и как следствие, предсказание характеристик параметров атаки). рисунок 5a показывает, что наибольшую сложностью для модели представляет задача различения ботов низкого и среднего качества; тем не менее, боты высокого качества все еще можно отличить от остальных.

В-четвертых, наиболее заметная корреляция (приблизительно 0,5) наблюдается между «Доверием» и TPR модели, как показано на рисунке 5b и в таблице 4, что говорит о том, что чем сложнее человеку распознать бота, тем сложнее его распознать и обученной модели. Тип продавца также оказывает влияние (корреляция примерно −0,2), что указывает на то, что боты, купленные на бирже, обнаруживаются несколько хуже. Корреляция с ценой, качеством и скоростью бота считается незначительной.

Таким образом, обученная модель подходит для обнаружения ботов, а эффективность обнаружения слабо коррелирует с основными характеристиками ботов – ценой, качеством, скоростью и типом продавца. Данные результаты указывают на то, что модели, обученные с использованием предложенного подхода, способны не только точно обнаружить значительную часть ботов, но и предсказать их характеристики, тем самым составив профиль атаки: определить скорости атаки ботов, связанные с атакой затраты, качество ботов, степень доверия пользователей к данным ботам и тип продавца. Более того, экспериментальная оценка указывает на то, что эффективность детектора не сильно зависит от типов ботов – модель может успешно идентифицировать большинство разновидностей ботов независимо от их параметров. Тем не менее, показатели обнаружения для некоторых наборов ботов остаются низкими, как показано на рисунке 4b, что намекает на наличие неучтенных «скрытых» параметров, которые могут влиять на эффективность детектора.

8. Заключение

В данной работе представлен подход к обнаружению ботов в социальной сети ВКонтакте и оценке их параметров. Использование общепринятых методов машинного обучения в сочетании с построением признаков на основе графов, текста и распределений, а также формированием наборов данных на основе метода покупки, позволило достигнуть высокой степени эффективности детектирования ботов и предсказания их метрик. Проведенные эксперименты указывают, что модели, построенные с использованием предложенного подхода, могут достаточно точно идентифицировать ботов даже на несбалансированных данных, а также обнаруживать большинство разновидностей ботов и предсказывать их метрики. Более того, в экспериментах была обнаружена только слабая корреляция эффективности обнаружения с основными характеристиками ботов, что указывает на то, что модель может успешно идентифицировать большинство разновидностей ботов независимо от их параметров. Сочетание предложенного подхода к обнаружению ботов с предсказанием метрик способен потенциально улучшить процессы выбора контрмер и противодействия атакам в социальных сетях, предоставляя защитным механизмам качественные характеристики атакующего и его возможностей.

Список используемых источников

1. Cresci S. A decade of social bot detection // Communications of the ACM. 2020. Vol. 63. Iss. 10. PP. 72–83. DOI:10.1145/3409116

2. Samoilenko S.A., Suvorova I. Artificial intelligence and deepfakes in strategic deception campaigns: The US and Russian experiences // In: The Palgrave Handbook of Malicious Use of AI and Psychological Security. Cham: Springer International Publishing, 2023. PP. 507–529. DOI:10.1007/978-3-031-22552-9_19

3. Yang K., Menczer F. Anatomy of an AI-powered malicious social botnet // arXiv preprint arXiv:2307.16336.2023. DOI:10.48550/arXiv.2307.16336

4. Gilani Z., Farahbakhsh R., Tyson G., Wang L., Crowcroft J. Of bots and humans (on twitter) // Proceedings of the International Conference on Advances in Social Networks Analysis and Mining. (New York, USA, 31 July 2017). Association for Computing Machinery, 2017. DOI:10.1145/3110025.3110090

5. Orabi M., Mouheb D., Al Aghbari Z., Kamel I. Detection of bots in social media: a systematic review // Information Processing and Management. 2020. Vol. 57. Iss. 4. P. 102250. DOI:10.1016/j.ipm.2020.102250

6. Коломеец М.В., Чечулин А.А. Метрики вредоносных социальных ботов // Труды учебных заведений связи. 2023. Т. 9. № 1. С. 94–104. DOI:10.31854/1813-324X-2023-9-1-94-104. EDN:HEFHFR

7. Zegzhda P.D., Malyshev E.V., Pavlenko E.Y. The use of an artificial neural network to detect automatically managed accounts in social networks // Automatic Control and Computer Sciences. 2017. Vol. 51. Iss. 8. PP. 874–880. DOI:10.3103/S0146411617080296. EDN:UYCEUW

8. Samokhvalov D.I. Machine learning-based malicious users' detection in the VKontakte social network // Proceedings of the Institute for System Programming of the RAS. 2020. Vol. 32. Iss. 3. PP. 109–117. DOI:10.15514/ISPRAS-2020-32(3)-10

9. Kaveeva A.D., Gurin K.E. Artificial VKontakte profiles and their impact on the social network of users // Journal of Sociology and Social Anthropology. 2018. Vol. 21. Iss. 2. PP. 214–231. DOI:10.31119/jssa.2018.21.2.8. EDN:XZOGHB

10. Kolomeets M., Tushkanova O., Levshun D., Chechulin A. Camouflaged bot detection using the friend list // Proceedings of the 29th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP, Valladolid, Spain, 10–12 March 2021). IEEE, 2021. PP. 253–259. DOI:10.1109/PDP52278.2021.00048. EDN:ZDXFHS

11. Skorniakov K., Turdakov D., Zhabotinsky A. Make Social Networks Clean Again: Graph Embedding and Stacking Classifiers for Bot Detection // Proceedings of the Workshops, co-located with 27th ACM International Conference on Information and Knowledge Management (CIKM, Torino, Italy, 22 October 2018). ISP RAS? 2019. Vol. 2482. EDN:IHZECD

12. Kolomeets M. MKMETRIC2022 – dataset with VKontakte bot identifiers and their metrics // guardeec/datasets. 2022. URL: https://github.com/guardeec/datasets#mkmetric2022 (Accessed 20.04.2024)

13. Zhou Z., Guan H., Bhat M., Hsu J. Detecting Fake News with NLP: Challenges and Possible Directions. 2018. URL: https://meghu2791.github.io/Fake_News_Detection.pdf (Accessed 20.04.2024)

References:

1. Cresci S. A decade of social bot detection. *Communications of the ACM*. 2020; 63(10):72–83. DOI:10.1145/3409116

2. Samoilenko S.A., Suvorova I. Artificial intelligence and deepfakes in strategic deception campaigns: The US and Russian experiences. In: *The Palgrave Handbook of Malicious Use of AI and Psychological Security.* Cham: Springer International Publishing; 2023. PP. 507–529. DOI:10.1007/978-3-031-22552-9_19

3. Yang K., Menczer F. Anatomy of an AI-powered malicious social botnet. *arXiv preprint arXiv:2307.16336.* 2023. DOI:10.48550/arXiv.2307.16336

4. Gilani Z., Farahbakhsh R., Tyson G., Wang L., Crowcroft J. Of bots and humans (on twitter). *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, 31 July 2017, New York, USA.* Association for Computing Machinery; 2017. DOI:10.1145/3110025.3110090

5. Orabi M., Mouheb D., Al Aghbari Z., Kamel I. Detection of bots in social media: a systematic review. *Information Processing and Management*. 2020;57(4):102250. DOI:10.1016/j.ipm.2020.102250

6. Kolomeets M., Chechulin A. Properties of Malicious Social Bots. *Proceedings of Telecommunication Universities.* 2023;9(1): 94–104. DOI:10.31854/1813-324X-2023-9-1-94-104. EDN:HEFHFR

7. Zegzhda P. D., Malyshev E. V., Pavlenko E. Y. The use of an artificial neural network to detect automatically managed accounts in social networks. *Automatic Control and Computer Sciences.* 2017;51(8):874–880. DOI:10.3103/S014641161708 0296. EDN:UYCEUW

8. Samokhvalov D.I. Machine learning-based malicious users' detection in the VKontakte social network. *Proceedings of the Institute for System Programming of the RAS*. 2020;32(3):109–117. DOI:10.15514/ISPRAS-2020-32(3)-10

9. Kaveeva A.D., Gurin K.E. Artificial VKontakte profiles and their impact on the social network of users // Journal of Sociology and Social Anthropology. 2018;21(2):214–231. DOI:10.31119/jssa.2018.21.2.8. EDN:XZOGHB

10. Kolomeets M., Tushkanova O., Levshun D., Chechulin A. Camouflaged bot detection using the friend list. *Proceedings of the 29th* Euromicro *International Conference on Parallel, Distributed, and Network-Based Processing, PDP, 10–12 March 2021, Valladolid, Spain.* IEEE; 2021. DOI:10.1109/PDP52278.2021.00048. EDN:ZDXFHS

11. Skorniakov K., Turdakov D., Zhabotinsky A. Make Social Networks Clean Again: Graph Embedding and Stacking Classifiers for Bot Detection. *Proceedings of the Workshops, co-located with 27th ACM International Conference on Information and Knowledge Management, CIKM, 22 October 2018, Torino, Italy, vol.2482,.* ISP RAS; 2019. EDN:IHZECD

12. Kolomeets M. MKMETRIC2022 – dataset with VKontakte bot identifiers and their metrics. *guardeec/datasets.* 2022. URL: https://github.com/guardeec/datasets#mkmetric2022 [Accessed 20 April 2024]

13. Zhou Z., Guan H, Bhat M., Hsu J. Detecting Fake News with NLP: Challenges and Possible Directions. 2018. URL: https://meghu2791.github.io/Fake_News_Detection.pdf [Accessed 20.04.2024]

Статья поступила в редакцию 29.01.2024; одобрена после рецензирования 26.03.2024; принята к публикации 08.04.2024.
The article was submitted 29.01.2024; approved after reviewing 26.03.2024; accepted for publication 08.04.2024.

Информация об авторах:

ЧЕЧУЛИН Андрей Алексеевич
кандидат технических наук, доцент, ведущий научный сотрудник лаборатории проблем компьютерной безопасности Санкт-Петербургского Федерального исследовательского центра Российской академии наук, доцент кафедры защищенных систем связи Санкт-Петербургского государственного университета телекоммуникаций им. проф. М.А. Бонч-Бруевича
https://orcid.org/0000-0001-7056-6972

КОЛОМЕЕЦ Максим Вадимович
научный сотрудник школы вычислительной техники Ньюкаслского университета
https://orcid.org/0000-0002-7873-2733