УДК 004.045

ОЦЕНКА ДОВЕРИЯ К РАЗЛИЧНЫМ ИНФОРМАЦИОННЫМ ИСТОЧНИКАМ И МЕТОДЫ РАБОТЫ С НИМИ В ЭПОХУ BIG DATA

Н.А. Калабин
Студент 1 курса специальности «Информатика и технологии программирования»
5203330@gmail.com

А.О. Чмутов
Студент 1 курса специальности «Информатика и технологии программирования»

В.Д. Владымцев
Ассистент кафедры информатики, инженер-программист ОИАСУ ЦИИР БГУИР
v.vladymtsev@bsuir.by

Н.А. Калабин
Студент 1-го курса БГУИР ФКСИС, по специальности ИИТП

А.О. Чмутов
Студент 1-го курса БГУИР ФКСИС, по специальности ИИТП

В.Д. Владымцев
Ассистент кафедры информатики, инженер-программист ОИАСУ ЦИИР БГУИР

Аннотация. Получать информацию стало проще, но многие источники теперь не отличаются качеством, что означает необходимость проведения надежных исследований. Дополнительная проблема заключается в том, что любой человек может распространять дезинформацию, что усложняет задачу отличить правду от лжи. Анализ больших данных может помочь в определении эффективных источников и методов получения информации, что впоследствии может привести к разработке более эффективных и надежных стратегий использования этих ресурсов.

Ключевые слова: Big Data, Большие Данные, Источники информации, Дезинформация, Доверие к источникам информации, Влияние.

Введение.

С приходом века высоких технологий получить доступ к информации стало как никогда просто, информация буквально окружает нас. Источники информации представлены в огромном количестве форм, таких как люди, документы, книги, веб-сайты, видео и различные другие носители знаний. Каждый из них предоставляет данные о конкретной теме или предмете. Сложно переоценить ценность информации. В умелых руках информация используется как инструмент войны: с её помощью свергают режимы, уничтожают человеческие судьбы, лоббируют нужные решения.

Дезинформация и роль аналитики больших данных в борьбе с ней.

Каналы средств массовой информации принято делить на две группы: печатные СМИ (газеты, журналы, книги и другие задокументированные формы) и электронные, которые, в свою очередь, включают как «традиционные» вещательные (радио и телевидение), так и «новые» медиа (социальные сети) [1], а также мы выделяем нейронные сети.

Традиционные источники информации, хоть и потеряли сегодня свою былую популярность [2], по-прежнему оказывают значительное влияние на восприятие, мнения и убеждения людей [1]. Однако, они имеют несколько проблем:

Традиционные источники часто основаны на доминирующих культурных нормах и ценностях, которые могут исключать или маргинализировать голоса, не соответствующие

этим нормам. Более того, традиционные источники информации подвержены цензуре, особенно в авторитарных режимах, что может ограничить доступ к информации, которая бросает вызов статусу-кво.

Кроме того, интерпретация традиционных источников информации часто зависит от индивидуальных предубеждений и мировоззрения читателя. Такая субъективность может привести к неправильному толкованию или искажению информации, увековечиванию дезинформации и укреплению общественных стереотипов.

Для контекстуализации этих вопросов стоит рассмотреть пример исторического текста, такого как "Одиссея" Гомера. Хотя "Одиссея" считается классическим произведением литературы, она также является продуктом своего времени, отражая доминирующие культурные нормы и ценности Древней Греции. Однако ее темы героизма и путешествия к самопознанию продолжают находить отклик у современных читателей. Поэтому традиционные источники информации следует анализировать критически, осознавая их контекст и ограниченность их перспектив.

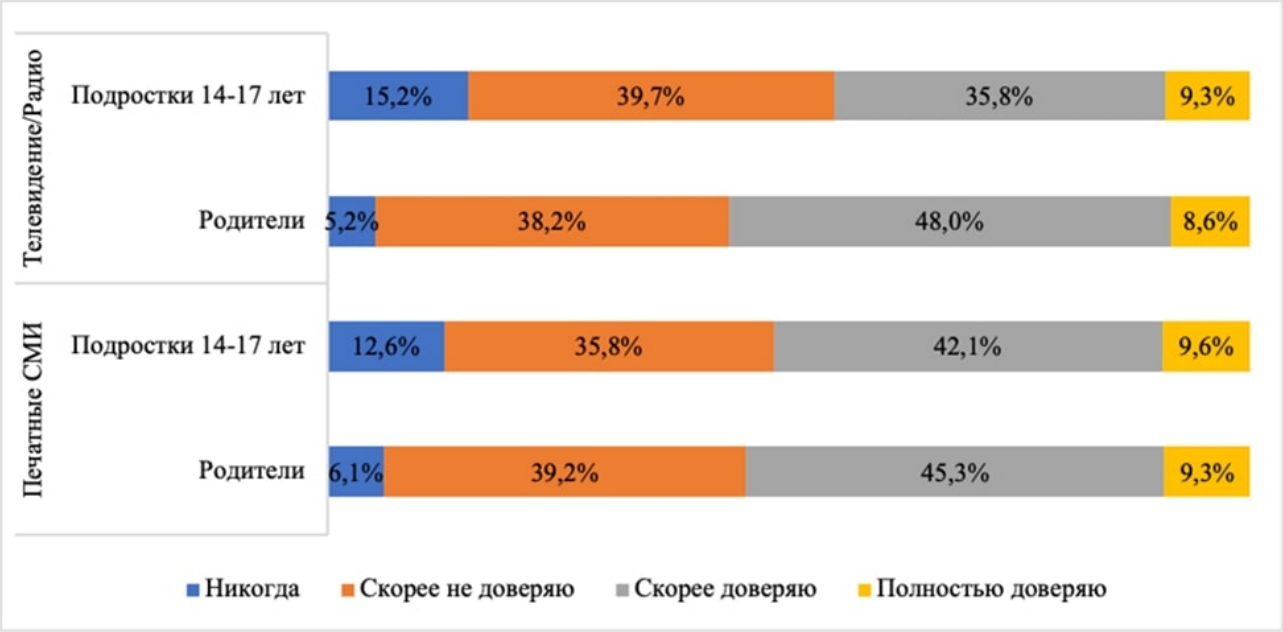

Рисунок 1. Доверие к информации из онлайн-источников у подростков и родителей, %. [1, с14]

Новые медиа сегодня преобладают над традиционными [2], они охватывают гигантскую аудиторию в реальном времени и делают информацию общедоступной. Доступность информации снизила ее качество, что привело к необходимости проводить собственные мини-исследования для получения достоверных сведений. Мы должны точно определить первоисточник и отличить надежные источники от ненадежных, чтобы получить достоверную информацию. Это занимает время и отнимает много сил, особенно у людей, которые родились до появления интернета. При этом нет гарантий, что вы найдете верную информацию. Эта проблема усугубляется тем, что любой человек, независимо от аккредитации или образования, теперь может выступать в качестве источника информации, т. к. увеличивается роль социальных сетей: в американском исследовании каждый второй взрослый называл их для себя постоянным источником новостной информации [3]. К сожалению, наказания за фабрикацию и распространение фальшивых историй незначительны по сравнению с огромным негативным воздействием, которое может

оказать дезинформация. В 2013 г. Всемирный экономический форум включил массовую цифровую дезинформацию в число наиболее серьезных глобальных рисков [4].

В силу того, какую роль сегодня играют социальные сети, они привлекают к себе много внимания, исследователи и практики стремятся понять динамику онлайн коммуникации, однако сталкиваются со множеством проблем с анализом данных:

Одной из проблем при анализе информации является качество самих данных. Платформы ежедневно генерируют большое количество информации, которая зачастую не релевантна, неточна или нерепрезентативна [5].

Однако данные социальных сетей не отражают мнение или поведение пользователей. Влияние оказывают социальные нормы, давление со стороны или самопрезентация. Необходимо применять строгие методы и техники для фильтрации, проверки и подтверждения данных [6].

Еще одной проблемой является сбор данных. Социальные сети часто накладывают ограничения на доступ к информации. Например, ограничивают количество запросов, временные рамки или объем, которые можно получить через API [5].

Кроме того, платформы социальных сетей могут менять свою политику, алгоритмы или функции без предварительного уведомления, что оказывает влияние на последовательность и сопоставимость.

Платформы не предоставляют достаточной информации о местоположении или структуре сети. Поэтому исследователи и аналитики нередко должны знать об ограничениях и проблемах, приводящих к поиску альтернативных источников.

Третья проблема — интерпретация данных. Данные часто являются сложными и многомерными, включая различные типы и форматы контента. Они часто неоднозначны и зависят от контекста, что требует глубокого понимания лингвистических, культурных и ситуационных аспектов коммуникации. Например, значение хэштега или эмодзи может меняться в зависимости от пользователя или платформы. Необходимо применять соответствующие методы и инструменты для анализа данных таким образом, чтобы отобразить богатство и разнообразие, а также учесть их неопределенность и изменчивость.

Четвертая проблема анализа информации из социальных сетей — этика данных. Данные социальных сетей включают персональную информацию о личности, предпочтениях, мнениях и поведении пользователей, что поднимает такие этические вопросы, как неприкосновенность частной жизни, согласие, анонимность, конфиденциальность и право собственности. Более того, данные социальных сетей могут иметь последствия для репутации, безопасности, прав и благополучия пользователей. Поэтому исследователям и аналитикам необходимо следовать этическим принципам и рекомендациям, чтобы защитить достоинства заинтересованных сторон.

Анализ информации из социальных сетей является сложным процессом, требующим тщательного рассмотрения различных аспектов. Эти критерии можно рассматривать и как возможности для разработки новых и инновационных подходов, повышающих ценность и влияние аналитики социальных сетей.

Социологические исследования показали, что подростки в возрасте 14–17 лет имеют высокий уровень доверия к социальным сетям, как источникам информации, так каждый второй сообщает о среднем уровне доверия и 11% полностью доверяют ленте в социальных сетях [1, с.14]. Это чрезмерное доверие молодого поколения часто приводит к тому, что они верят ложной информации и делятся ею без предварительной проверки.

По сравнению с детьми, у родителей уровень доверия к онлайн-ресурсам ниже, а к традиционным медиа — выше, что подтверждается и в зарубежных исследованиях [7].

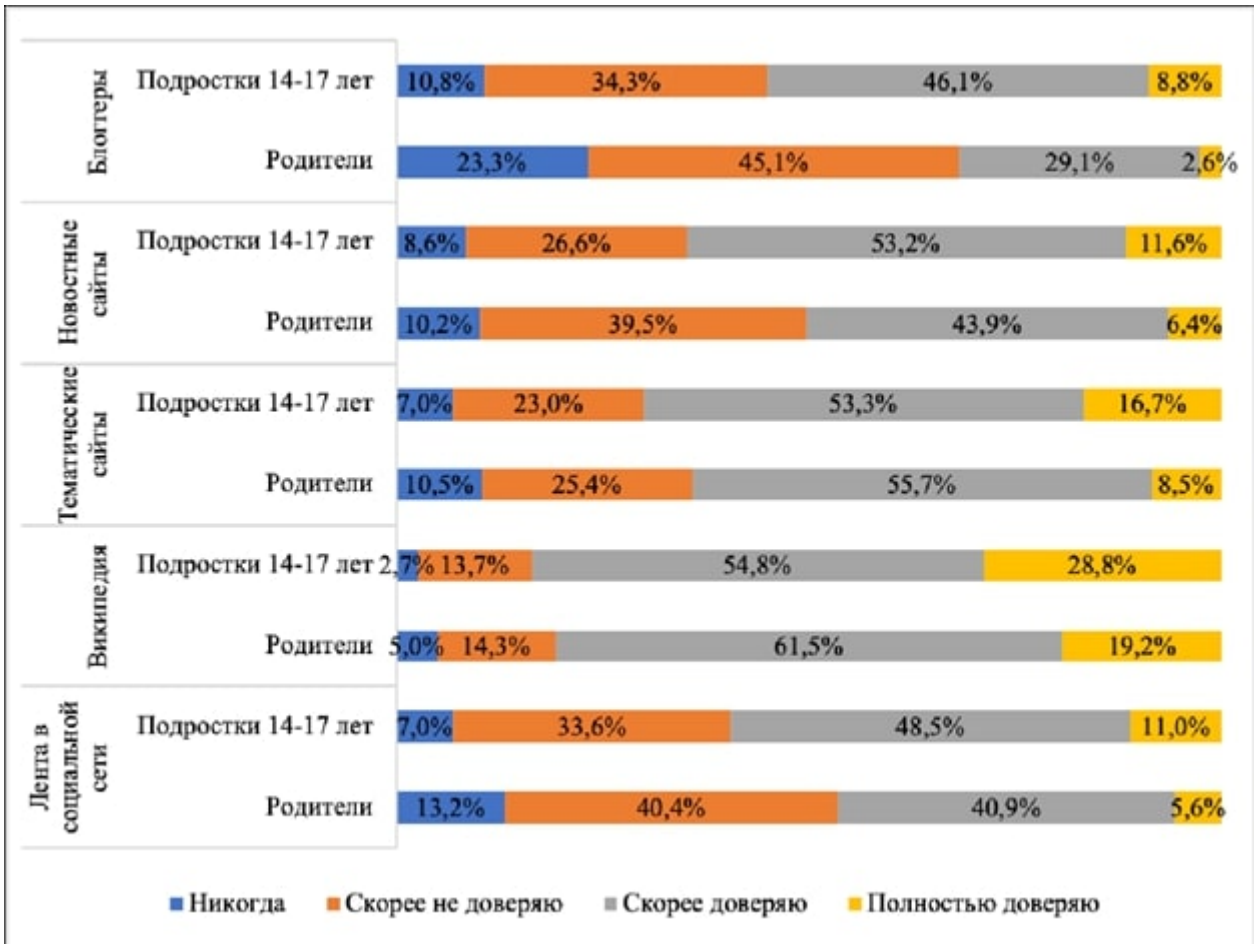

Рисунок 2. Доверие к информации из онлайн-источников у подростков и родителей, %. [1, с14]

Невозможно не учесть, как источник информации, появившиеся совсем недавно, нейронные сети, основанные на методе глубокого обучения, они позволяют человеку за считанные секунды получить конкретный ответ на свой вопрос. На данный момент, Chat GPT, в основе которого лежит мультимодальная большая языковая модель GPT-4 [8], является наиболее продвинутой нейронной сетью, она проходит собеседования на ведущие позиции в ИТ компании, сдаёт экзамены на уровне 10% лучших выпускников ВУЗов [9]. Кроме GPT-4 существует множество других языковых моделей, таких как BERT, T-NLG, Switch-C, ELMo и др. Но они все уступают GPT-4 на данный момент как по количеству параметров, так и по качеству и точности ответов.

При вопросе языковая модель анализирует текст, чтобы воспринять его и ответить соответствующим образом. Её понимание языка и контекста позволяет создать ответ, предсказывая идеальную последовательность слов, извлеченных из массы текстовых данных, которые она изучила. По сути, языковая модель генерирует ответ, предсказывая, каким будет следующее наиболее вероятное слово(а), основываясь на вводимом тексте и контексте разговора. Она также может генерировать несколько возможных ответов и выбирать наиболее подходящий на основе различных факторов, таких как релевантность, согласованность и грамматическая корректность.

Одной из основных проблем языковых моделей является то, что они очень сильно зависят от той информации, которую им предоставили в момент обучения, ведь именно на её основании строятся ответы [10]. Например, исследователи обнаружили, что BERT

ассоциирует фразы, в которых упоминаются инвалиды, с более негативными словами настроения, и что насилие с применением оружия, бездомность и наркомания чрезмерно представлены в текстах, обсуждающих психические заболевания [11]. Аналогично, модели типа GPT-3, обученные на не менее чем 570 ГБ данных, полученных в основном из Common Crawl 16 [12], могут генерировать предложения с высокими оценками токсичности, даже когда им предлагаются нетоксичные предложения [13]. А при изучении обучающих данных для GPT-217 обнаружили 272 тыс. документов с ненадежных новостных сайтов и 63 тыс. документов из запрещенных веток Реддита. Из-за всего вышеперечисленного ответы на некоторые вопросы могут быть некорректными или ошибочными.

Второй проблемой можно назвать контроль, хоть в этом направлении и предпринимаются серьезные усилия [9], но языковые модели всё ещё можно взломать и получить от них запрещенную информацию.

Третья проблема — это сложность интерпретации. Языковые модели могут быть сложны в интерпретации из-за их нелинейности и сложных внутренних структур, что затрудняет объяснение принятых решений и выводов.

GPT-4 и другие языковые модели на основе глубокого обучения — это невероятно мощные и сложные инструменты, которые отвечают на запросы пользователей, используя огромные объемы текстовых данных. Несмотря на их сложность, они не лишены ограничений. Например, GPT-3 может столкнуться с проблемами контроля, интерпретации и ошибками из-за зависимости от обучающих данных. Важно подходить к этим моделям с осторожностью, учитывая их ограничения и недостатки.

При углублении в научные исследования, посвященные доверию к различным источникам информации и эффективным механизмам преодоления трудностей стоит изучить применение аналитики больших данных.

Анализ огромных массивов данных позволяет выявить охват и эффективность источников информации в различных областях жизни, включая науку, бизнес, медицину, политику и многое другое — потенциальная золотая жила для понимания социальных тенденций. Обработка больших данных позволяет выявить модели потребления и предпочтения аудитории, основанные на различных переменных, таких как возраст, пол, образование или местоположение.

Существует несколько подходов к изучению использования источников информации в различных областях и выявлению наиболее эффективных методов для достижения конкретных целей. Таких как:

Исследование пользователей — это распространенный метод, используемый для сбора информации среди пользователей, которые ищут решения различных проблем в своей профессиональной или личной жизни. Чтобы получить эти данные, исследователи проводят ряд исследований, которые могут включать в себя опросы, интервью, фокус-группы и тестирование продуктов. Конечный результат исследования играет ключевую роль в определении наиболее эффективных источников информации в конкретном сценарии.

Анализ данных — это альтернативный метод, который предполагает глубокое изучение использования различных источников информации. Можно анализировать данные, связанные с использованием различных источников информации в рабочей среде, таких как журналы посещений веб-сайтов, статистику использования информационных систем и другие метрики. Анализ данных может помочь определить наиболее эффективные источники информации в конкретной ситуации.

Исследование лучших практик: одним из эффективных методов изучения источников информации является исследование передового опыта в различных областях деятельности. Изучая, как организации и компании эффективно используют свои информационные

ресурсы, вы можете получить ценные знания, которые можно применить в собственной работе.

Экспертное мнение: консультации с экспертами в вашей области могут дать ценные сведения об эффективных источниках информации для достижения ваших целей и задач. Делясь своим опытом и советами в конкретных ситуациях, эксперты могут предложить уникальные рекомендации по источникам информации, которые могут быть непредсказуемыми, но плодотворными для расширения ваших знаний и навыков.

Определение наиболее эффективных источников информации в конкретном сценарии имеет ключевое значение и может быть осуществлено с помощью различных методов. Для получения всестороннего анализа рекомендуется применять комбинацию различных подходов, имеющихся в вашем распоряжении. Исследование больших данных может дать представление о частоте и эффективности различных методов и практик, связанных с источниками информации, в различных секторах.

Заключение.

С помощью анализа больших данных можно выявить тенденции в использовании источников информации и методов работы с ними, определить их эффективность в различных сферах бизнеса, что приведет к разработке более мощных и надежных методов работы с этими ресурсами.

Список литературы

[1] Солдатова Г.У., Чигарькова С.В., Илюхина С.Н. Медиапотребление подростков и родителей: источники информации и доверие к ним // Психологические исследования. 2021. Т. 14, № 80. С. 4. URL: http://psystudy.ru/index.php/num/2021v14n80/1955-soldatova80.html с-14.

[2]Медиапотребление в России-2020, URL: https://oohmag.ru/wp-content/uploads/2020/11/mediapotreblenie-v-rossii-2020.pdf

[3] Shearer E., Mitchell A. News Use Across Social Media Platforms in 2020 // Pew Research Center, 2021. URL: https://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020/

[4] Howell L. Digital wildfires in a hyperconnected world // WEF Report 2013. World Economic Forum. Eight editions. 2013. 25 p.

[5] Stieglitz Stefan. The Adoption of Social Media Analytics for Crisis Management — Challenges and Opportunities, 2018. URL: https://www.researchgate.net/publication/325416290_The_Adoption_of_Social_Media_Analytics_for_Crisis_Management_-_Challenges_and_Opportunities

[6] Weiguo Fan, Michael D. Gordon. Unveiling the Power of Social Media Analytics, 2014. URL: https://www.researchgate.net/publication/259148570_The_Power_of_Social_Media_Analytics

[7] Edelman Trust Barometer. Global Report, 2019, 65 p.

[8] GPT-4 — Википедия (wikipedia.org), URL: https://ru.wikipedia.org/wiki/GPT-4

[9] GPT-4 Technical Report, url: https://arxiv.org/abs/2303.08774

[10] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic
Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

[11] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In Proceedings of the 58th Annual Meeting of the Associ-ation for Computational Linguistics. Association for Computational Linguistics, Online, 5491–5501. https://doi.org/10.18653/v1/2020.acl-main.487

[12]https://commoncrawl.org/

[13] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[14] Pause Giant AI Experiments: An Open Letter, URL: https://futureoflife.org/open-letter/pause-giant-ai-experiments/

ASSESSMENT OF TRUST IN VARIOUS INFORMATION SOURCES AND METHODS OF WORKING WITH THEM IN THE ERA OF BIG DATA

N.A. Kalabin
BSUIR student,
system programmer engineer

A.O. Chmutau
BSUIR student,
system programmer engineer

V.D. Vladymtsev
Assistant of the Department of Computer Science, Software Engineer of DIACS CIIR BSUIR

Belarusian state university of informatics and radioelectronics, Republic of Belarus
E-mail: 5203330@gmail.com, v.vladymtsev@bsuir.by

Annotation. It has become easier to get information, but many sources are no longer of high quality, which means that reliable research is needed. An additional problem is that anyone can spread disinformation, which complicates the task of distinguishing truth from lies. Big data analysis can help in identifying effective sources and methods of obtaining information, which can subsequently lead to the development of more effective and reliable strategies for using these resources.

Keywords: Big Data, Information sources, Disinformation, Trust in information sources, Influence.