

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 3](#_Toc10902071)

[1 Теоретические основы исследования технологий Big Data 5](#_Toc10902072)

[1.1 Разработка технологии Big Data в специальной литературе 5](#_Toc10902073)

[1.2 Позиционирование технологии Big Data в прошлом и настоящем 11](#_Toc10902074)

[2 Технологии Big Data: контент обобществления коммуникаций 19](#_Toc10902075)

[2.1 Анализ использования технологий Big Data в социальных сетях 19](#_Toc10902076)

[2.2 Проблемы, возможности и перспективы использования технологии Big Data в социальных сетях 23](#_Toc10902077)

[ЗАКЛЮЧЕНИЕ 26](#_Toc10902078)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 27](#_Toc10902079)

# ВВЕДЕНИЕ

Актуальность использования технологии Big Data достаточно велика, так как в данный момент она является одним из ключевых драйверов развития информационных технологий. Это направление, относительно новое для российского бизнеса, получило широкое распространение в западных странах. Связано это с тем, что в эпоху информационных технологий, особенно после бума социальных сетей, по каждому пользователю интернета стало накапливаться значительное количество информации, что в конечном счете дало развитие направлению Big Data.

Цель данной работы заключается в изучении теоретических исследований технологии Big Data, а также способов её использования в социальных сетях.

Основные задачи данной работы:

* выявить состояние разработок, связанных с большими данными, в академической среде;
* изучить, как менялись взгляды на большие данные с течением времени
* продемонстрировать влияние информационных технологий на коммуникации в обществе;
* проанализировать способы использования больших данных в социальных сетях;
* выявить проблемы и перспективы применения технологий, связанных с большими данными, в социальных сетях.

Объектом исследования данной работы являются практика применения технологии Big Data в социальных сетях. Предмет исследования – совокупность социально-экономических отношений, выражающихся в виде научных статей, публикаций в СМИ, отчётов консалтинговых компаний, которые формируют представление о сущности и перспективах использования технологии Big Data, в том числе в социальных сетях.

Теоретическую базу исследования составляют позиционирование и представление технологии Big Data в научной среде.

Методологическую базу исследования составляют анализ и синтез, метод научных абстракций и моделирование, индукция и библиографический анализ.

Эмпирической основой курсовой работы являются аналитические обзоры, личные наблюдения и исследования.

Теоретическая значимость данной работы заключается объединении и систематизации накопленной информации по данной теме. Рассмотрение в данной конкретных примеров использования технологии Big Data представляет практическую значимость данной работы.

Полученные результаты могут быть использованы в дальнейшем в практической деятельности рекламных агентств.

Курсовая работа состоит из введения, двух глав, заключения и списка использованных источников. **1 Теоретические основы исследования технологий Big Data**

## 1.1 Разработка технологии Big Data в специальной литературе

Исследования, связанные с большими данными [1-12], являются актуальной областью разработок на сегодняшний день. Они изучают большие данные, а также способы выделения знаний из них. Они ведутся в рамках различных дисциплин и областей, таких как информационные науки, моделирование неопределённости, машинное обучение, статистическое обучение, распознавание образов, методы хранения данных, обработка сигналов и т. д. Исследования больших данных также имеют собственные проблемы и задачи. Проведём обзор ситуации в академической среде, охватывающей Big Data. Рассмотрим основные группы проблем исследований, связанных с большими данными.

В рамках первой группы проблем рассматриваются проблемы, связанные с хранением и анализом больших данных. Они изучаются исследователями Z. Huang, T. K. Das, P. M. Kumar, D. P. Acharjya и M. R. Patra.

Проблема хранения возникает в силу увеличения скорости создания новых данных в последние годы. Благодаря мобильным устройствам, интернету вещей, увеличению доступности интернета и прочим факторам, объём производимой информации растёт в геометрической прогрессии. В силу недостатка места для их хранения они либо удаляются, либо не записываются вовсе. В связи с этим, возрастает роль носителей информации и скорости её записи и чтения для доступности больших данных с целью их анализа. Несмотря на достижения в этой области, такие как, например, распространение твердотельных накопителей, необходимая производительность накопителей для обработки больших данных до сих пор не достигнута [1].

Кроме объёмов производимой информации, в последние годы также нарастает и их разнообразие, что значительно усложняет задачи анализа больших данных. Возникает необходимость сжатии или выборке обрабатываемых данных, поскольку существующее методы и алгоритмы не позволяют за приемлемое время произвести их анализ. Автоматизация этого процесса, в том числе при помощи машинного обучения, и является задачей, стоящей перед исследователями [2].

Последние технологии, такие как Hadoop и MapReduce, позволяют собирать большие объёмы полуупорядоченных и неупорядоченных данных за приемлемое время. Для получения возможности их дальнейшей обработки они нуждаются в упорядочивании. Разработка алгоритмов упорядочивания данных также является актуальной задачей [3].

В рамках второй группы проблем исследований рассматриваются проблемы выделения знаний из больших данных и вычислительная сложность обработки больших данных. Они изучаются исследователями Al-Jarrah O. Y., Yoo P. D., Muhaidat S. и Karagiannidis G. K.

Выделение и представление знаний из больших данных – главная задача их обработки. Она включает в себя несколько подзадач, таких как аутентификация, архивирование, управление, сохранение, поиск и представление знаний. Алгоритмы, которые используются для решения этих задач, основаны, по большей части, на теории нечётких множеств и нечёткой логики, которые в настоящее время активно развиваются.

Анализ больших данных может представлять высокую вычислительную сложность. Главной проблемой при анализе является устранение несоответствий и неопределённости, которые присутствуют в наборах данных. Несмотря на то, что попытки преодоления вычислительной сложности реализуются в большинстве случаев обработки наборов больших данных, единого метода, применимого ко всем случаям, до сих пор не существует. Имеющиеся инструменты анализа имеют слишком низкую производительность и не в состоянии эффективно справляться с несоответствиями, неопределённостью и вычислительной сложностью , которые возникают при обработке наборов больших данных. Разработки в этой области уже имеются, а также проводятся новые, преимущественно с использованием машинного обучения. Главной целью ставится минимизация вычислительной сложности [4], [5].

Третью группу проблем исследований составляют проблемы масштабируемости и визуализации больших данных. Они изучаются исследователем A. Jacobs. В последнее время исследования в области больших данных позволили добиться ускорения их обработки, на фоне увеличения производительности процессоров по закону Мура. Несмотря на это, объёмы больших данных растут гораздо быстрее, чем производительность процессоров. В связи с этим, ставится задача распараллеливания вычислений между разными процессорами, в том числе между различными ядрами одного и того же процессора. Методы и алгоритмы параллельных вычислений являются одной из областей исследования.

Цель визуализации набора больших данных – дать аналитикам адекватное представление о его свойствах, помочь правильно их интерпретировать. Визуализация позволяет превратить большой массив данных в графики или изображения, которые дадут аналитикам интуитивное представление об их содержании. Современные инструменты визуализации обладают неудовлетворительной производительностью, функционалом и временем отклика, что также является проблемой исследований.

Четвёртая группа проблем исследований представлена проблемами информационной безопасности, связанной с большими данными. Они изучаются исследователями Zhu H., Xu Z., Huang Y. В процессе анализа наборов больших данных из них добываются полезные знания. Разные организации имеют различную политику безопасности для защиты конфиденциальной информации, которая необходима в связи с высокими рисками, сопровождающими операции с большими данными [6]. Информационная безопасность становится проблемой при анализе больших данных. Она может быть обеспечена при помощи техник аутентификации, авторизации и шифрования. Меры безопасности больших данных существуют в условиях компьютерных сетей, разнообразия девайсов, отслеживанием безопасности в реальном времени и проверкой на отсутствие утечек информации. Несмотря на наличие разработок в этой области, меры по обеспечению безопасности больших данных всё ещё нуждаются в улучшении. Главная задача состоит в том, чтобы разработать многоуровневую систему безопасности, обеспечивающую полноценную защиту приватности при обработке больших данных.

Рассмотрим ещё несколько актуальных, но менее масштабных областей исследований, связанных с большими данными: интернет вещей, облачные вычисления и биоинспирированные вычисления.

Интернет вещей — концепция вычислительной сети физических предметов, оснащённых встроенными технологиями для взаимодействия друг с другом или с внешней средой, рассматривающая организацию таких сетей как явление, способное перестроить экономические и общественные процессы, исключающее из части действий и операций необходимость участия человека [7]. В процессе своей работы «вещи» генерируют непрерывный поток данных. Технологии больших данных могут быть применены для его обработки, что позволить выделять из них знания и совершенствовать управление этими данными, что позволит добиться большей эффективности работы «вещей». Большого прогресса в этой области достигли исследователи N. Mishra, C. Lin и H. Chang [8]. На рисунке 1 представлен цикл использования знаний из больших данных.

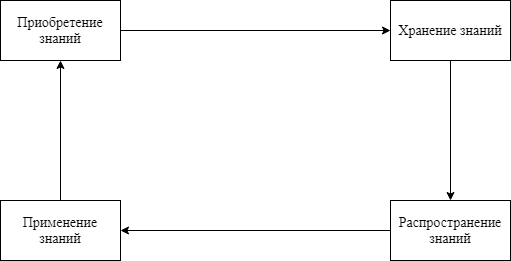


Рисунок 1 – «Цикл использования знаний из больших данных»

На первом этапе, приобретении знаний, знания выделяются с использованием различных вычислительных методов. На втором этапе организуется хранение полученных знаний в специально спроектированной для этой задачи системе. На третьем этапе из хранящихся знаний выделяется значимая для принятия какого-либо решения информация, а на четвёртом она используется для принятия решения, как человеком, так и компьютером. Принятое может отразиться на следующей итерации, обуславливая то, по каким схемам будет происходить следующее выделение знаний.

Достижения в области визуализации сделали вычисления на суперкомпьютерах более доступными. В то время как пользователь видит лишь результат визуализации, вычисления проводятся при помощи удалённой от него компьютерной инфраструктуры, вычислительная мощность которых многократно превосходит мощность персональных компьютеров. Данная технология носит название облачных вычислений и является ключевой в обработке больших данных. Спрос на неё обусловлен нерентабельностью владения инфраструктуры с большой вычислительной мощностью на постоянной основе, в то время как результаты её работы могут быть необходимы. В настоящее время существуют исследования в этой области, с упором на управление и хранение данных [9], [10].

Ещё одним преимуществом облачных вычислений является возможность облачного хранения данных. Если данные будут храниться в другом месте, возникают трудности с их загрузкой и выгрузкой для осуществления вычислений, что является проблемой в силу их большого объёма.

С использованием облачных технологий сопряжена ещё одна проблема – безопасность данных. Так как данные и результаты их обработки должны быть переданы по сети, возникает опасность их перехвата. Решение всех вышеперечисленных проблем позволит вывести обработку больших данных на новый уровень.

Биоинспирированные вычисления – это технология построения алгоритмов, которые копируют принципы работы биологических систем. Главными преимуществами этих алгоритмов являются простота и высокая скорость нахождения оптимального решения [11]. Перспективы применения биоинспирированных вычислений обсуждались исследователями C. Shi, Y. Shi, Q. Qin и R. Bai [12]. Согласно их выводам, биоинспирированные вычисления имеют большой потенциал для решения задач, связанных с большими данными.

Объём данных, генерируемых из различных источников, удваивается каждые два года. Их обработка открывает широкие перспективы для получения полезной информации, однако требует соответствующих методов. Преобразование больших данных в знания – комплексная и сложная задача. Данные могут включать в себя неопределённость, методы для преодоления которой уже существуют (нечёткая логика, нейронные сети), но далеки от совершенства. Данные зачастую полезны не в полном объёме, что требует методов их отбора для конкретного исследования или задачи. Все эти методы требуют высоких вычислительных мощностей и объёмов памяти, что также является проблемой. Мы рассмотрели основные направления исследований, связанных с большими данными. Рассмотрим теперь, как к феномену больших данных относились в прошлом и настоящем.

## 1.2 Позиционирование технологии Big Data в прошлом и настоящем

Понимание феномена больших данных менялось от момента его возникновения и до наших дней. Рассмотрим ключевые события, которые характеризуют смену и углубление взглядов на большие данные.

Сам термин Big Data был впервые введен в 1997 году Michael Cox и David Ellsworth на 8-й конференции IEEE по визуализации. Они назвали проблемой больших данных нехватку емкости основной памяти, локального и удаленного диска для выполнения виртуализации [13]. А в 1998 году руководитель исследовательских работ в SGI John Mashey на конференции USENIX использовал термин Big Data в его современном виде [14].

Понимание возможностей больших данных появилось немного позже. Так, в ноябре 2000 года на восьмом всемирном конгрессе эконометрического сообщества Francis Diebold представил доклад под названием «Big Data Dynamic Factor Models for Macroeconomic Measurement and Fore-casting», в котором утверждал следующее: «В последнее время наука столкнулась с феноменом больших данных и извлекла из них выгоду. Большие данные – это рост количества (а иногда и качества) доступных и потенциально важных данных, который является следствием высоких достижений в области записи и хранения информации» [15].

6 февраля 2001 года Douglas Laney из Meta Group (входит в состав Gartner) издал документ, описывающий основные проблемные зоны, связанные с повышенными требованиями к центральным хранилищам данных на фоне бурного роста e-commerce, а также делающий прогноз на изменение стратегии IT-компаний в отношении подходов к построению архитектуры решений, связанных с хранением и обработкой информации.

Было выделено три важнейших направления, на которых стоит сосредоточиться для решения вопросов управления данными: Volume (объёмы данных), Velocity (скорость накопления и обработки данных) и Variety (разнообразие источников и типов данных). Позже эти понятия стали основой для описательной модели больших данных под названием 3V (VVV).

Нужно учесть, что эти аспекты обсуждались без отсылки к понятию больших данных, но эти параметры описали основные принципы того, что сегодня называется Big Data [16].

Внимание широкой общественности к большим данным было привлечено в июне 2008 года, когда в журнале «Wired» вышла статья Chris Anderson «The end of theory: the data deluge makes the scientific method obsolete» [17]. В ней утверждалось, что всё возрастающий объём данных позволяет науке строить прогнозы, не формируя для этого теории. Знание о корреляции между величинами может быть достаточным для принятия решения.

Широкое введение термина «большие данные» в научной среде связывают с Clifford Lynch, редактором журнала Nature, подготовившим к 3 сентября 2008 года специальный выпуск с темой «Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объёмами данных?», в котором были собраны материалы о феномене взрывного роста объёмов и многообразия обрабатываемых данных и технологических перспективах в парадигме вероятного скачка от количества к качеству.

В декабре 2008 года Randal E. Bryant, Randy H. Katz и Edward D. Lazowska опубликовали статью «Big-data computing: creating revolutionary breakthroughs in commerce, science and society». В ней утверждалось: «Подобно поисковым системам, которые изменили степень доступности информации, обработка больших данных может изменить деятельность компаний, научных исследований, медицину и оборону… Обработка больших данных, вероятно, является самым большим достижением в области вычислительной техники за последнее десятилетие. Мы только начали понимать, какой потенциал несут большие данные во всех сферах жизни. Государственные инвестиции могут значительно ускорить исследования в этой области».

Формировались взгляды на большие данные и в бизнесе. Так, в мае 2009 года исследовательская и консалтинговая компания Gartner опубликовала документ, который прогнозировал рост данных, хранимых предприятиями, на 650% в следующие 5 лет. В нём также утверждалось, что если будет найден способ извлекать из этих данных полезные знания, то это может привести к революции в бизнесе.

В 2011 году McKinsey Global Institute опубликовал отчёт, в котором был приведён анализ технологий Big Data. Большие данные рассматриваются сразу в трех плоскостях – роста объемов, роста скорости обмена данными и увеличения информационного разнообразия. Согласно отчёту, существует пять основных способов, которыми использование больших данных может создать ценность. Во-первых, большие данные могут создать значительную ценность, делая информацию прозрачной и пригодной для использования с гораздо большей скоростью. Во-вторых, поскольку организации создают и хранят больше данных о транзакциях в цифровом виде, они могут собирать более точную и подробную информацию о производительности по всему, начиная с инвентарных списков продуктов и заканчивая больничными днями, и, следовательно, демонстрируют адаптивность и повышают производительность. Ведущие компании используют сбор и анализ данных для проведения контролируемых экспериментов с целью принятия лучших управленческих решений; другие используют данные для прогнозирования, чтобы своевременно вносить изменения в свой бизнес. В-третьих, большие данные обеспечивают более узкую сегментацию клиентов и, следовательно, более точные рекламные предложения. В-четвертых, аналитика больших данных может существенно улучшить процесс принятия решений. Наконец, большие данные могут быть использованы для улучшения разработки продуктов и услуг следующего поколения. Например, производители используют данные, полученные от датчиков, встроенных в продукты, для создания инновационных предложений послепродажного обслуживания, таких как проактивное обслуживание (профилактические меры, которые направлены на предотвращение сбоев и поломок). Также утверждается, что большие данные станут ключевой основой для конкуренции, которая послужит росту продуктивности и роста для отдельных фирм. В то время как использование больших данных будет иметь значение для разных отраслей экономики, некоторые отрасли настроены на большую прибыль. IT, а также финансы и страхование готовы существенно выиграть от использования больших данных [18].

В мае 2012 года Danah Boyd и Kate Crawford опубликовали статью «Critical Questions for Big Data». Они определяют большие данные как «культурный, технологический и научный феномен, основанный на взаимодействии технологии (увеличение вычислительной мощности и алгоритмической точности для сбора, анализа, связывания и сравнения наборов данных) и анализа (использование наборов больших данных для выявления закономерностей с целью предъявления экономических, социальных, технических и юридический требований)» [19].

6 октября 2015 года стало известно об исключении из отчета Gartner «Hype Cycle 2015» сведений о больших данных. Свое решение аналитики компании объяснили тем, что в состав понятия «большие данные» входит большое количество технологий, активно применяющихся на предприятиях, они частично относятся к другим популярным сферам и тенденциям и стали повседневным рабочим инструментом., а в науке исследования, связанные с большими данными рассредоточились в рамках прикладных дисциплин и областей [20].

## 1.3 Социализация информационных технологий

Социальная коммуникация — это:

* передача информации, идей, эмоций посредством знаков, символов;
* процесс, который связывает отдельные части соц. систем друг с другом;
* механизм, посредством которого реализуется власть (власть, как попытка определить поведение другого человека).

Имеются различные подходы к типологии коммуникации.

По масштабности процесса коммуникации и массовости вовлекаемых в него лиц различают массовую (на уровне социальной системы), среднего уровня (ограниченную в масштабах социальных групп и организаций - внутриорганизационные коммуникации) и локальную (внутрисемейная, триадная и др.), внутригрупповую (взаимодействие не выходит за рамки определенной группы), межгрупповую (между различными группами или большой группы с ее достаточно самостоятельными подгруппами.

По отношению к каждому из внешних субъектов межгрупповых взаимосвязей коммуникации подразделяют на: внешние коммуникации, межличностные (между отдельными людьми - диадные) и внутриличностные (общение с самим собой).

Коммуникации по способу установления и поддержания контакта подразделяются на непосредственные (прямые), опосредованные (дистанционные). Коммуникация непосредственная - коммуникация, осуществляемая напрямую с использованием вербальных и невербальных средств в пределах визуального восприятия (например, беседа, публичное выступление). Опосредованная коммуникация - взаимодействие, осуществляемая через посредника. Коммуникация при этом осуществляется как посредством физических лиц - посредников, так и различных средств коммуникации, включая технические (средства массовой информации, такие технические средства как рупор, громкоговоритель, телефон, радио, видеосвязь, компьютерные коммуникационные сети локального или глобального характера и другие информационные технологии).

Информационная технология - совокупность методов, производственных и программно-технологических средств, объединенных в технологическую цепочку, обеспечивающую сбор, хранение, обработку, вывод и распространение информации.

Ускоряющаяся информатизация является предсказуемым, наиболее видимым сейчас вектором воздействия новых технологий на социальную структуру общества. Такое воздействие можно разбить на две группы.

Непосредственное воздействие новых информационных технологий на сознание человека – в первую очередь, за счет значительного расширения его когнитивных и коммуникативных возможностей. При этом мало предсказуемыми являются ответы на вопрос о влиянии (прямом или косвенном) распространения новых информационных технологий на динамику изменения системы ценностей.

Воздействие информационных технологий на социальную организацию: использование новых средств коммуникации принципиально снижает издержки передачи информации, упрощается и стремительно ускоряется процесс создания географически распределенных социальных групп (сетей), характеризующихся преобладанием неиерархических «горизонтальных» коммуникаций, возможностью общения «всех со всеми».

Потребность человека в информации и необходимость её поиска в информационных ресурсах общества предполагают коммуникацию между человеком и информационными службами и создают ряд социально-психологических коллизий.

Согласно теории социальной коммуникации, коммуникация — это взаимодействие двух субъектов, где один субъект – коммуникант, отправитель, передатчик; второй – адресат, приемник, реципиент. Взаимодействие этих субъектов опосредуется передаваемым объектом.

Если применить это определение коммуникации к информационному обслуживанию, то становится ясно, что последнее – социальная коммуникация, поскольку в нем наличествует два субъекта (информационный работник и пользователь) и опосредующие их взаимодействие объекты – запрос, документ. Можно выделить в этом процессе еще каналы связи, барьеры коммуникации, информационную среду, в которой осуществляется взаимодействие.

Информационная и социально-коммуникационная стороны информационного обслуживания обусловлены связью социальной информации и социальной коммуникации: коммуникация предполагает передачу информации, исходя из этого, информация не может стать доступной без коммуникации.

Одной из особенностей информационного обслуживания как коммуникации является то, что роли субъектов в ходе этой деятельности меняются: при предъявлении пользователем запроса информационному работнику пользователь является коммуникантом, отправителем, а информационный работник – адресатом, приемником; при выполнении запроса и передаче информационным работником документа или информации пользователю первый становится коммуникантом, отправителем, второй – адресатом, приемником.

В ходе получения услуг между пользователем и информационной службой (информационным работником) осуществляется коммуникация.

Коммуникация пользователей и информационной службы осуществляется в устной и письменной формах, непосредственно при посещении информационного учреждения и через каналы связи с использованием технических средств.

В связи с внедрением новых информационных технологий говорят об удаленном (виртуальном) пользователе [21].

# **2 Технологии Big Data: контент обобществления коммуникаций**

## 2.1 Анализ использования технологий Big Data в социальных сетях

Приступим к рассмотрению использования Big Data в социальных сетях на конкретных примерах.

Крупнейшая социальная сеть в мире – Facebook, использует большие данные для распознавания изображений, технологии, которая позволяет компьютеру идентифицировать объект на изображении или видео, основываясь на результатах анализа миллионов похожих изображений. Распознавание изображений позволяет Facebook предлагать пользователям отмечать на своих фотографиях других людей, распознавая их на фотографии. Также, технология корректирует ленту новостей пользователя. Например, если друзья пользователя часто ставят «лайк» изображениям с собаками, то пользователь начнёт видеть похожие изображения в своей ленте чаще.

Большое количество данных о пользователях также позволяет Facebook эффективно размещать таргетинговую рекламу. Социальная сеть помогает бизнесу в поиске потенциальных клиентов, основываясь на данных об их интересах и склонностях [22]. Рекламодатели имеют возможность очень точно конфигурировать целевую аудиторию, а также получить от Facebook предложения по её расширению [23].

Когда речь заходит о данных пользователей, возникают вопросы приватности. В контексте рекламы Facebook решила эту проблему при помощи своей технологии «Тематических данных». Тематические данные (Topic Data) — это специальная технология Facebook, которая показывает маркетологам отзывы людей о брендах и событиях, но таким образом, чтобы вся информация осталась конфиденциальной. В свою очередь, маркетологи уже используют полученные данные для изменения стратегии размещения бренда как на Facebook, так и на других платформах. Ранее такие данные можно было получить только от специализированных компаний, однако выборка была небольшой, а определение демографии — вовсе почти невозможно. С помощью Topic Data соцсеть смогла сгруппировать данные для маркетологов по узким направлениям, убрав при этом личную информацию пользователей [24].

На свой 10-летний юбилей Facebook предложил пользователям посмотреть и поделиться видео, в котором показаны их самые популярные посты и действия со дня регистрации до настоящего времени. Такие видео называются флешбеками и состоят из фотографий, комментариев и постов, которые зациклены в слайдшоу под приятную музыку. Такие же видеозаписи Facebook автоматически создает на годовщины вашей дружбы с другими пользователями, с которыми вы хоть как-то были связаны внутри соцсети (есть общие фотографии или посты с тегами), а также в день вашего рождения [25].

Занимая шестое место в списке самых популярных социальных сетей во всем мире, Instagram стремится сделать платформу максимально толерантной. Для этой цели они используют DeepText, технологию Facebook, которая имеет возможность почти с человеческой точностью анализировать несколько тысяч постов пользователей в секунду.

Для определения тона и намерения комментария, его правильной интерпретации и различения между оскорблением и конструктивной критикой, технология Instagram также изучает контекстуальное значение смежных комментариев.

Кроме того, DeepText помогает в распознавании спама. Система определяет фейковые аккаунты и удаляет спам в комментариях под постами пользователей. Эта функция пока доступна не на всех языках и ещё находится в разработке.

Перед введением в использование система проанализировала и сгруппировала более двух миллионов комментариев по группам, таким как «оскорбление», «расизм», «сексуальные домогательства». Если пользователь не хочет видеть комментарии подобного рода, он может включить соответствующую функцию в настройках своей учётной записи.

Twitter, ещё одна крупная социальная сеть, использует машинное обучения для эффективной обрезки изображений. В ленте пользователя изображения не показываются в полном виде, а лишь в виде фрагментов. Совершенствуя свои алгоритмы, Twitter обучает свою систему выбирать наиболее интересные для пользователя фрагменты – лица, текст, животных.

Поскольку нейронные сети в рамках подобных задач работают слишком медленно, чтобы осуществлять работу в реальном времени, Twitter использует два метода. Первый их них, «дистилляция знаний», используется для обучения меньшей сети для имитации работы более большой и осуществления предсказания, основываясь на наборе изображений и сторонних важных данных. Второй из них, сокращение Фишера, используется для удаления лишних фрагментов при одновременном снижении затрат вычислительных мощностей.

Такое сочетание методов позволяет ускорить работу нейронной сети и обрабатывать изображения, как только они были загружены – в 10 раз быстрее, чем при стандартном подходе. Пример использования технологии приведён на рисунке 2.



Рисунок 2 – Пример использования технологии обрезки изображений

Для Reddit — центра интернет-новостей, изображений, историй, мемов и видео расширенный поиск является приоритетом. По этой причине социальная сеть внедряет лучшие технологии для предоставления высококачественного контента.

Стремясь сделать свой поиск релевантным, быстрым, и легко масштабируемым с ростом контента, Reddit внедряет платформу на основе искусственного интеллекта под названием «Fusion». Это помогает компании решить задачу индексации, объединяя данные из разных источников в единое представление. Reddit не только индексирует новые посты, но и обновляет их релевантность в реальном времени, основываясь на количестве комментариев и голосов. Внедрение этой технологии дало следующие результаты:

* количество индексированных постов увеличилось на 33 %,
* период реиндексации всего сайта уменьшился с одиннадцати до пяти часов,
* уровень ошибок при поиске сократился вдвое, при этом 99 % результатов поиска были обработаны менее чем за 500 мс,
* количество компьютеров, вовлечённых в поисковые операции, сократилось с двухсот до тридцати [25].

## 2.2 Проблемы, возможности и перспективы использования технологии Big Data в социальных сетях

Технология Big Data, несмотря на свою полезность в социальных сетях, имеет и ряд проблем. Также, в дополнение к имеющимся возможностям, она имеет потенциальные возможности и перспективы. Рассмотрим наиболее важные их них.

Главной проблемой использования больших данных в социальных сетях является приватность. Социальные сети собирают множество информации о своих пользователях, включая переписку, данные о расположении, фотографии, история активности. С использованием технологий анализа больших данных, из этой информации можно извлечь множество дополнительных сведений, распространение которых может быть нежелательно. Например, исследователи из Стэнфордского университета продемонстрировали возможность определения сексуальной ориентации человека, анализируя его фотографию. По одной фотографии нейросеть верно определяет ориентацию в 81% случаев для мужчин и в 74% для женщин. Если же исследователи давали нейросети пять фотографий одного человека, то точность вырастала до 91% и 83% соответственно. В связи с этой проблемой социальные сети совершенствуют политику конфиденциальности. Facebook, в частности, не передаёт рекламодателям исходные данные о пользователях, а лишь результаты их анализа и обработки.

Ещё одной проблемой является технический аспект хранения и обработки больших данных. Имеющиеся продукты не являются в полной мере удовлетворяющими потребности компаний. Facebook, например, в работе с Big Data полагается в большей степени на одну технологию — Hadoop. Она представляет собой масштабируемую среду с открытым исходным кодом, которая использует большое количество серверов для работы. Анализ данных при помощи этой технологии начинается со считывания 300 петабайт данных — информация по конкретному запросу изымается из базы и помещается в специальную таблицу, обработкой которой занимается специальный отдел Facebook. Чтобы уменьшить временные и денежные затраты на анализ, компания занялась разработкой собственного гигантского хранилища данных, что позволит сократить расходы по многим параметрам [27].

Рассмотрим теперь потенциальные возможности использования больших данных в социальных сетях:

* Прогнозирование оттока пользователей. Предсказывание удаления профиля пользователями. Это поможет поддерживать отношения с пользователями с высоким уровнем риска;
* Прогноз ценности пользователей как покупателей. Предсказывание того, сколько денег потратит конкретный пользователь на ту или иную группу товаров. Анализируя их поведение, искусственный интеллект может определить их привычки и побудить их тратить деньги, показывая таргетированную рекламу;
* Анализ настроений постов и комментариев. Может описать отношение аудитории к конкретному продукту или бренду;
* Персонализация контента. Алгоритмы машинного обучения, определяющие предпочтения каждого пользователя, а затем рекомендующие соответствующий контент для них;
* Прогноз популярности контента. Популярность контента может быть связана со спросом на определённые товары, что можно использовать при подборе рекламы [28].

В целом потенциальные перспективы использования больших данных сводятся к одной цели – повышение эффективности рекламных предложений. Реализуется это либо напрямую, предложением наиболее востребованной рекламы, либо косвенно, повышением привлекательности использования социальной сети, что увеличивает количество просматриваемых рекламных предложений.

# ЗАКЛЮЧЕНИЕ

По результатам проделанной работы можно сделать следующие выводы:

* имеющиеся трудности, с которыми сталкиваются разработчики и исследователи технологий больших данных, являются разнообразными и комплексными, однако их решение способно привести к значительным усовершенствованиям в этой области;
* c момента возникновения феномена больших данных его понимание претерпевало изменения, сам феномен вызвал широкомасштабные обсуждения как в научной среде, так и в бизнесе, а на настоящий момент он утратил свою новизну и является рядовым инструментом как учёных, так и коммерческих предприятий;
* информационные технологии оказали значительное влияние на человеческие коммуникации;
* применение больших данных в социальных сетях подчинено одной цели – повышению эффективности рекламы;
* практика применения больших данных в социальных сетях будет расширяться и изменяться.

Поставленные задачи были решены частично. Для полноценного их решения требуется более глубокий анализ как научных публикаций, так и публикаций в СМИ.

Полученные результаты могут быть использованы для первичного ознакомления с феноменом больших данных в социальных сетях.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Huang Z. A fast clustering algorithm to cluster very large categorical data sets in data mining // SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. – 1997.

2 Das T. K., Kumar P. M. Big data analytics: A framework for unstructured data analysis // International Journal of Engineering and Technology. – №5(1). – 2013. – pp.153-156.

3 Das T. K., Acharjya D. P., Patra M. R. Opinion mining about a product by analyzing public tweets in twitter // International Conference on Computer Communication and Informatics. – 2014.

4 Al-Jarrah O. Y., Yoo P. D., Muhaidat S., Karagiannidis G. K. Efficient Machine Learning for Big Data: A Review // Big Data Research. – №61.– 2015. – pp. 87-93.

5 Changwon. Y, Ramirez L., Liuzzi J., Big Data Analysis Using Modern Statistical and Machine Learning Methods in Medicine // International Neurourology Journal. – 85. – 2014, pp. 50.

6 Zhu H., Xu Z., Huang Y. Research on the security technology of big data information // International Conference on Information Technology and Management Innovation. – №55. – 2015. – pp.1041-1044.

7 Ashton. L That ‘Internet of Things’ Thing. In the real world, things matter more than ideas // RFID Journal. – №53. – 2009.

8 Mishra N., Lin C., Chang H. A cognitive adopted framework for IOT big data management and knowledge discovery prospective // International Journal of Distributed Sensor Networks. – №66. – 2015. – pp. 1-13.

9 Assuno M. D., Calheiros R. N., Bianchi S., Netto M. a. S., Buyya R. Big data computing and clouds: Trends and future directions // Journal of Parallel and Distributed Computing. – №79. – 2015. – pp. 3-15.

10 Hashem I. A. T., Yaqoob I., Badrul Anuar N., Mokhtar S., Gani A., Ullah Khan S. The rise of big data on cloud computing: Review and open research issues // Information Systems. – №47. – 2014. – pp. 98-115.

11 Wang L. and Shen J. Bioinspired cost-effective access to big data // International Symposium for Next Generation Infrastructure. – №43. – 2013. – pp. 1-7.

12 Shi C., Shi Y., Qin Q., Bai R., Yin H., Tang K., Gao Y., Klawonn F. Swarm Intelligence in Big Data Analytics // Intelligent Data Engineering and Automated Learning. – №35. – 2013. – pp. 417-426.

13 Application-Controlled Demand Paging for Out-of-Core Visualization. – URL: https://www.evl.uic.edu/cavern/rg/20040525\_renambot/Viz/parallel\_volviz/paging\_outofcore\_viz97.pdf (дата обращения 6.06.19).

14 Big Data and the Next Wave of InfraStress. – URL: http://static.usenix.org/event/usenix99/invited\_talks/mashey.pdf (дата обращения 6.06.19).

15 Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting. – URL: https://www.sas.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF (дата обращения 6.06.19).

16 История больших данных. Часть 1. – URL: https://www.computerra.ru/234239/istoriya-bolshih-dannyh-big-data-chast-1/ (дата обращения 6.06.19).

17 The end of theory: the data deluge makes the scientific method obsolete. – URL: https://www.wired.com/2008/06/pb-theory/ (дата обращения 6.06.19).

18 Big data: The next frontier for innovation, competition, and productivity. – URL: https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI\_big\_data\_full\_report.ashx (дата обращения 6.06.19).

19 Boyd. D, Crawford. K. Critical Questions for Big Data // Information, Communication & Society. – №15. – 2012. – pp. 662–679.

20 Большие данные (Big Data). – URL: http://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:%D0%91%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D0%B5\_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5\_(Big\_Data) (дата обращения 6.06.19).

21 Роль информационных технологий в социальной коммуникации. – URL: https://students-library.com/library/read/28600-rol-informacionnyh-tehnologij-v-socialnoj-kommunikacii (дата обращения 6.06.19).

22 Want to Use Big Data? Why Not Start Via Google, Facebook, Amazon, (Etc.). – URL: https://www.forbes.com/sites/bernardmarr/2017/08/14/want-to-use-big-data-why-not-start-via-google-facebook-amazon-etc/ (дата обращения 6.06.19).

23 How big data is affecting social media metrics and Facebook ad strategies. – URL: https://www.smartdatacollective.com/how-big-data-affecting-social-media-metrics-and-facebook-ad-strategies/ (дата обращения 6.06.19).

24 Topic Data: Learn What Matters to Your Audience. – URL: https://www.facebook.com/business/news/topic-data (дата обращения 6.06.19).

25 Facebook: A Decade of Big Data. – URL: https://www.wired.com/insights/2014/03/facebook-decade-big-data/ (дата обращения 6.06.19).

26 How Social Media Giants Leverage Big Data and ML to Serve Users Better. -URL: https://socialmediaweek.org/blog/2019/01/how-social-media-giants-leverage-big-data-and-ml-to-serve-users-better/ (дата обращения 6.06.19).

27 Как Facebook работает с Big Data. – URL: https://ain.ua/2017/12/09/kak-facebook-rabotaet-s-big-data/ (дата обращения 6.06.19).

28 Machine Learning in Social Media: 6 ways how ML is monetizing big data. – URL: https://addepto.com/machine-learning-social-media-monetizing-data/ (дата обращения 6.06.19).