МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ КАЗАХСТАН
Некоммерческое акционерное общество
АЛМАТИНСКИЙ УНИВЕРСИТЕТ ЭНЕРГЕТИКИ И СВЯЗИ
имени Гумарбека Даукеева

Кафедра «Телекоммуникационные сети и системы»
Специальность: 6М071900 «Радиотехника, электроника и телекоммуникации»

ДОПУЩЕН К ЗАЩИТЕ
Зав. кафедрой
PhD, доцент Темырканова Э.К.
(ученая степень, звание, ФИО)

_____
(подпись)
«_____»_____2020 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ
пояснительная записка

на тему: «Диффузия информации в распределенных системах»

Магистрант: Замахов А.В. _____ группа МРЭТн-18-2
(Ф.И.О.) (подпись)
Руководитель: Ph.D., доцент _____ Семенякин Н.В.
(ученая степень, звание) (подпись) (Ф.И.О.)

Рецензент_____ _____ _____
(ученая степень, звание) (подпись) (Ф.И.О.)

Консультант по ВТ Ph.D., доцент _____ Семенякин Н.В.
(ученая степень, звание) (подпись) (Ф.И.О.)

Нормоконтроль: Ph.D., доцент _____ Темырканова Э.К.
(ученая степень, звание) (подпись) (Ф.И.О.)

Алматы 2020

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ КАЗАХСТАН
Некоммерческое акционерное общество
АЛМАТИНСКИЙ УНИВЕРСИТЕТ ЭНЕРГЕТИКИ И СВЯЗИ
имени Гумарбека Даукеева

Институт Космической Инженерии и Телекоммуникаций

Специальность: 6М071900 «Радиотехника, электроника и телекоммуникации»

Кафедра: «Телекоммуникационные сети и системы»

ЗАДАНИЕ
на выполнение магистерской диссертации

Магистранту Замахову Александру Владиславовичу
(фамилия, имя, отчество)

Тема диссертации «Диффузия информации в распределенных системах»

Утверждена Ученым советом университета №122 от «25» 10.2018

Срок сдачи законченной диссертации «25» мая 2020г.

Цель исследования является построение общей методологии анализа диффузии информации в распределенных сетях, на основе изучения распространения информации по определенной тематике в ряде отечественных СМИ, а также построение на основе такого анализа картины распространения информации в ретроспективе.

Перечень подлежащих разработке в магистерской диссертации вопросов или краткое содержание магистерской диссертации:

1. Теоретические основы процесса распространения информации в распределенных сетях
2. Расчет размера сети распространения информации и вычисление количества агентов диффузии информации.
3. Определение метрик СМИ влияющих на диффузии в распределенных системах
4. Формирование списка 25 самых востребованных СМИ в РК и сбор датасета по размещенному контенту
5. Анализ процента взаимосвязи между распространителями информации и формирования графа диффузии в распределенной системе.

Перечень графического материала (с точным указанием обязательных чертежей)

Рисунок 2.4 – Общий размер сети распространения информации

Рисунок 2.9 – Корреляция между количеством посетителей и количеством просмотров

Рисунок 2.15 – Анализ текстов статей с различных информационных порталов на сходство при 1 слове в шингле

Рисунок 3.1 – Процесс диффузии информации между СМИ во время телевизионного обращения главы Республики Казахстан

Рекомендуемая основная литература

1. Губанов Д.А., Информационные процессы в социальных сетях (на примере сети Хабрахабр) // Интернет-конференция по проблемам управления.: ИПУ РАН. – 2010.

2. Mandelli, A., Accoto, C., & Mari, A.. Social media metrics: Practices of measuring brand equity and reputation in online social collectives. Proceedings of the 6th International Conference Thought Leaders in Brand Management, Lugano, Switzerland, 2010

Г Р А Ф И К
подготовки магистерской диссертации

Наименование разделов, перечень разрабатываемых вопросов	Сроки представления научному руководителю	Примечание
1. Информационный обзор согласно теме	05.10.2019	
2. Организация сбора первоначального датасета для проведения исследования	16.01.2020	
3. Анализ ключевых метрик 41 СМИ Республики Казахстан, которые оказывают влияние на распространение информации	28.01.2020	
4. Анализ процента взаимосвязи между размещенным контентом в ряде СМИ с помощью алгоритма шинглов	04.04.2020	
5. Формирование графа диффузии информации изучаемой новости	02.05.2020	

Дата выдачи задания_30 сентября 2018г._____

Заведующий кафедрой _____ (Темырканова Э.К.)
(подпись) (Ф.И.О.)

Научный
руководитель диссертации _____ (Семенякин Н.В.)
(подпись) (Ф.И.О.)

Задание принял к исполнению
магистрант _____ (Замахов А.В)
(подпись) (Ф.И.О.)

Аңдатпа

Бұл тезисте Қазақстан Республикасының аймағына байланысты жалпы ақпарат тарату желісінің көлеміне баға берілді, жалпы ақпараттық кеңістіктегі ресурстардың әрқайсысының үлесін анықтау үшін медиа-нарыққа талдау жасалды, әлеуетті неғұрлым ықпалды бұқаралық ақпарат құралдары анықталды, сондай-ақ оларды бағалау параметрлері анықталды. Көрсетілген бұқаралық ақпарат құралдарының жұмыс аймағымен байланысы болды, сонымен бірге нақты тақырыптарға сілтемелер келтірілген деректер жинақталды. Осының бәрі ақпарат көздерінің арасындағы байланысты анықтауға және таратылған жүйеде ақпараттың жалпы көрінісін қалыптастыруға мүмкіндік берді. Ақпаратты тарату графигін талдау нәтижесінде барлық таратылған желінің негізгі қатысушыларын анықтауға мүмкіндік берді.

Аннотация

В данной диссертационной работе была проведена оценка размера общей сети распространения информации в зависимости от региона РК, проведена аналитика всего рынка СМИ для определения доли каждого из ресурсов в общем информационном пространстве, выявлены потенциальные наиболее влиятельные СМИ, а также параметры их оценки. Указанные СМИ были соотнесены с регионом их работы, а также был сформирован набор данных с упоминаниями по определенной тематике. Все это позволило в конечном итоге определить взаимосвязи между источниками и сформировать общую картину распространения информации в распределенной системе. Анализ самого графа диффузии информации в качестве результата позволил определить главных участников во всей распределенной сети.

Abstract

In this dissertation work, was made an assessment of the size of the general information dissemination network depending on the region of the Republic of Kazakhstan, an analysis of the entire media market was carried out to determine the share of each of the resources in the overall information space, were identified the potential most influential media, as well as parameters for its assessment. The indicated media were correlated with the region of their work, and was formed a dataset with references on a specific topic. All this made it possible to ultimately determine the relationship between sources and form a general picture of the distribution of information in a distributed system. An analysis of the information diffusion graph itself as a result made it possible to identify the main participants in the entire distributed network.

Содержание

Введение

За последние пятнадцать лет развитие сети Интернет позволило миллиардам пользователей по всему миру без особых трудностей производить и потреблять контент. За счет участия в распределенных сетях стало возможным получать доступ к огромному числу источников информации. У каждого человека благодаря социальным сетям появился доступ к собственным СМИ, которые могут оказывать значительное влияние на процесс распространения информации. Однако роль традиционных СМИ тоже не снизилась, а наоборот возросла, так как получить к ним доступ стало еще проще за счет развития интернет покрытия в стране.

При этом с ростом количества информации, все важнее становится вопрос о ее качестве. Так как на население любой страны ведется постоянное информационное воздействие из различных источников, очень важно определить первоисточник информации и выяснить подлинность этой информации.

Фейковые новости представляют из себя намеренное распространение недостоверной информации в традиционных СМИ и социальных сетях с целью получения финансовой или политической выгоды за счет введения в заблуждение. Для генерации фейковых новостей используется множество подходов, таких как: провокационные заголовки, полностью недостоверные истории, воздействие на эмоции читателя – все это направлено на повышение количества аудитории и процента цитируемости. Причем распространение информации в данном случае происходит по принципу снежного кома: чем больше источников ссылается на недостоверный ресурс, тем выше вероятность появления нового источника распространения фейковой информации. Прибыль при этом формируется за счет доходов от рекламы или за счет финансирования кампании по распространению информации со стороны третьих лиц. Упрощение доступа к получению финансирования за счет рекламы, а также поляризация политического общества способствуют увеличению количества фейковой информации.

Анализ диффузии информации, а также воссоздание картины взаимодействия между источниками могут позволить определять первоисточники распространения недостоверных данных.

С помощью таких исследований можно выяснить, как идеи распространяются среди групп людей. Диффузия выходит за рамки двухэтапной теории потока, концентрируясь на условиях, которые увеличивают или уменьшают вероятность того, что инновация, новая идея, продукт или практика будут приняты членами данной культуры. В многоступенчатом распространении лидер мнений все еще оказывает большое влияние на поведение отдельных лиц, называемых последователями, но существуют и другие связующие звенья между средствами массовой информации и принятием решений аудиторией. Одним из посредников

является агент распространения информации, который поощряет лидера общественного мнения принять или отклонить новшество.

Информация не принимается всеми людьми в социальной системе одновременно. Вместо этого они, как правило, адаптируются во временной последовательности и могут быть классифицированы по категориям последователей в зависимости от того, сколько времени им потребуется, чтобы начать использовать новую идею. Практически говоря, агенту очень полезно иметь возможность определить, к какой категории принадлежат определенные лица, поскольку краткосрочная цель большинства агентов распространения информации состоит в том, чтобы способствовать внедрению инноваций. Принятие новой идеи обусловлено взаимодействием человека через межличностные сети. Если первоначальный инициатор инновации обсуждает его с двумя членами данной социальной системы, и эти два становятся последователями, которые передают инновацию двум партнерам и т. д. В данной диссертационной работе такими агентами распространения информации выступают СМИ.

Целями это работы является построение общей методологии анализа диффузии информации в распределенных сетях, на основе изучения распространения информации по определенной тематике в ряде отечественных СМИ, а также построение на основе такого анализа картины распространения информации в ретроспективе. Объектом при этом является инфоповод, который был одним из самых крупных в 2019 году, а также о котором имеется достаточно большое количество упоминаний и в СМИ – прекращение полномочий первого президента Республики Казахстан Назарбаева Н.А. Еще одним важным фактором в пользу этого инфоповода является то что, можно достоверно определить время начала распространения информации, а также выяснить главный первоисточник информации, так как он представляет из себя государственный ресурс.

Достижение этой цели будет происходить в несколько этапов: первоначально необходимо выявить что представляет из себя процесс распространения информации и какие существуют устойчивые паттерны. После чего будет проведен анализ влияния СМИ как агента распространения информации в распределенной сети на конкретного индивида. Далее будут рассмотрены принципы выявления взаимосвязей между источниками и анализа контента, на примере анализа фейковых новостей. Практический эксперимент будет представлять из себя тоже несколько основных фаз: определение размера общей сети, поиск потенциальных источников распространения информации, сбор датасета для анализа, проведение исследования на наличие взаимосвязей между источниками.

Результатом же этой диссертационной работы может выступить картина распространения информации среди СМИ, а также выявление ресурсов, оказавших наибольшее влияние на этот процесс.

1 Методология исследования диффузии информации

Вопрос распространения информации внутри сети, состоящей из представителей небольшой группы, впервые был освящен Марком Грановеттером в его основополагающей статье «Сила слабых связей» в 1973 году и был областью активных научных исследований в последние три десятилетия [1].

С того времени объем, направленность и способы распространения информации в распределенных сетях претерпели значительные изменения. Появились простые инструменты, с помощью которых возможно транслировать мнение одного человека на широкую аудиторию. Однако принципы, которые были заложены еще 40 лет назад еще актуальны, так как они основываются на психологических и физиологических особенностях общения между людьми.

Прежде чем погрузиться в процесс анализа, что такое диффузия информации, необходимо дать определение самому термину информация.

В свою очередь это тоже вызывает множество философских дискуссий на тему что можно считать информацией. Одним из вариантов определения является: информация — это знания, переданные или полученные относительно определенного факта или обстоятельства. Информация, знания, мудрость - это термины для человеческих приобретений через чтение, изучение и практический опыт.

Информация относится к изложенным, прочитанным или переданным фактам, которые могут быть неорганизованными и даже не связанными: собирать полезную информацию. Знание - это организованная совокупность информации или понимание и понимание, вытекающие из приобретения и упорядочения совокупности фактов: знания химии. Мудрость - это знание людей, жизни и поведения, причем факты настолько тщательно усвоены, что породили проницательность, суждение и понимание: использовать мудрость в обращении с людьми [2].

Информация может рассматриваться как разрешение неопределенности; это то, что отвечает на вопрос «что такое сущность» и, таким образом, определяет, как ее сущность, так и характер ее характеристик. Понятие информации имеет разные значения в разных контекстах. Таким образом, понятие становится связанным с понятиями ограничения, общения, контроля, данных, формы, образования, знаний, значения, понимания, умственных стимулов, паттерна, восприятия, репрезентации и энтропии [3].

Следовательно, сам термин информация частично и подразумевает обмен и распространение некими данными между участниками сети.

Распространение информации является обширной областью исследований и привлекает исследовательские интересы из многих областей, таких как физика, биология и т. д. Распространение инноваций между элементами сети является одной из первоначальных причин для изучения сетей. Однако похожие принципы применялись уже на протяжении веков,

когда ученые пытались понять, как распространяются заболевания среди людей.

Грановеттер предположил, что основной обмен информацией происходит по так называемым слабым связям - слабые связи в стиле «знакомства» между членами социальных сетей, в то время как прочные связи («стиль дружбы») отвечают за принятие решений, формирование и сохранение знаний, Теория слабых и прочных связей с небольшими изменениями и существенными дополнениями служит основой для современных теорий распространения информации. Грановеттер предположил, что что-то (информация, инновации) распространяется в социальных сетях без учета каких-либо конкретных механизмов. Информация может распространяется в виде сообщений, настенных сообщений и даже так называемых однобитных сигналов. Кайе Дж. [4] и Цзяо З. [5]. Голдер С. [6] провели первое опубликованное массивное исследование динамики сообщений в Facebook и обнаружили, что сообщения следуют обычным временным моделям, основанным на времени дня, дня недели и времени года. Источники и места назначения сообщений также не случайны, что свидетельствует о гомофильности процессов распространения.

Навигация в сложных сетях (не обязательно в массивных), т.е. эффективные механизмы поиска предполагаемых целей коммуникации, была смоделирована Богуной [7]. Они предложили концепцию скрытого метрического пространства, лежащего в основе сложной сети, с ее собственными координатами и расстояниями, которые служат руководством для принятия решений о маршрутизации. На наблюдаемом уровне навигационная способность выражается в безмасштабном (степенном законе) распределении степеней узлов и сильной кластеризации, оба свойства являются общими в сетях распространения информации, которые, таким образом, являются навигационными

Косинец [8], основываясь на работе Гибсона [9], заметил, что члены социальной сети, будь то онлайн или оффлайн, не постоянно общаются со своими соседями. Информация распространяется только в результате дискретных коммуникационных событий, и частоты этих событий оказывают сильное влияние на предпочтительные пути коммуникации. Эти пути образуют коммуникационную магистраль, которая динамически изменяется, чтобы отразить мгновенные колебания частот событий.

Таким образом мы можем заметить, что процесс изучения моделей диффузии информации является комплексным, и чаще всего в основе исследований лежит взаимодействие между отдельными людьми в сети. Однако в данной диссертационной работе будет рассмотрено более крупное взаимодействие, а именно: на уровне обмена информацией СМИ между собой.

Однако для того чтобы понять, как происходит само взаимодействие нужно определиться еще с несколькими понятиями. К ним относятся: источники информации, пути распространения, а также механизмы взаимодействия.

1.1 Основные источники, пути и механизмы диффузии информации

В ходе исследования, которое провел в 2004 году Груль Д. [10], до недавнего времени стоимость технической инфраструктуры, необходимой для охвата большого числа людей, выступала основным препятствием для тех, кто хотел распространять информацию, инновации или влияние через сообщество.

Сегодня, с распространением Интернета и массовых онлайновых социальных сетей и к интернету в целом, эта техническая проблема была в значительной степени устранена. По мере того, как все больше и больше людей получают доступ к интернету и становятся источниками информации, становится очевидно, что, если понять природу взаимодействия между людьми внутри сети, это позволит значительно расширить методологии, применяемые как в маркетинге, проектировании мобильных и веб приложение и в процессе информирования населения в целом.

Далее будут рассмотрены основные источники информации и характер взаимодействия между ними, для понимания структуры сети распространения информации.

1.1.1 Различие между типами связей участников сети распространения информации

Социальную сеть можно определить, как «социальную структуру, состоящую из отдельных лиц (или организаций), называемых узлами, которые связаны (связаны) одним или несколькими конкретными типами взаимозависимости, такими как дружба, родство, общие интересы, финансовый обмен, неприязнь или отношения верований, знаний или престижа. Социальные связи в сети служат двойному назначению: они указывают уровень близости между узлами и обеспечивают каналы связи. Социальные сети являются полем изучения анализа социальных сетей, который, стал ключевой техникой в современной социологии.

Он также получил значительное признание в области антропологии, биологии, коммуникационных исследований, экономики, географии, информатики, организационных исследований, социальной психологии и социолингвистики и стал популярной темой спекуляций и различных исследований.

Анализ социальной сети начинается с построения ее формального графа G (V, E). Граф G состоит из вершин V (также известных как узлы) и ребер E (также называемых связями, связями или дугами. Узел в социальном графе представляет участника (отдельное лицо или организацию), а связь представляет отношение или взаимозависимость между любыми двумя членами (Рисунок 1.1).

Граф состоит из 20 вершин (узлов), 25 сильных связей и 4 слабых связей, а также двух кластеров.

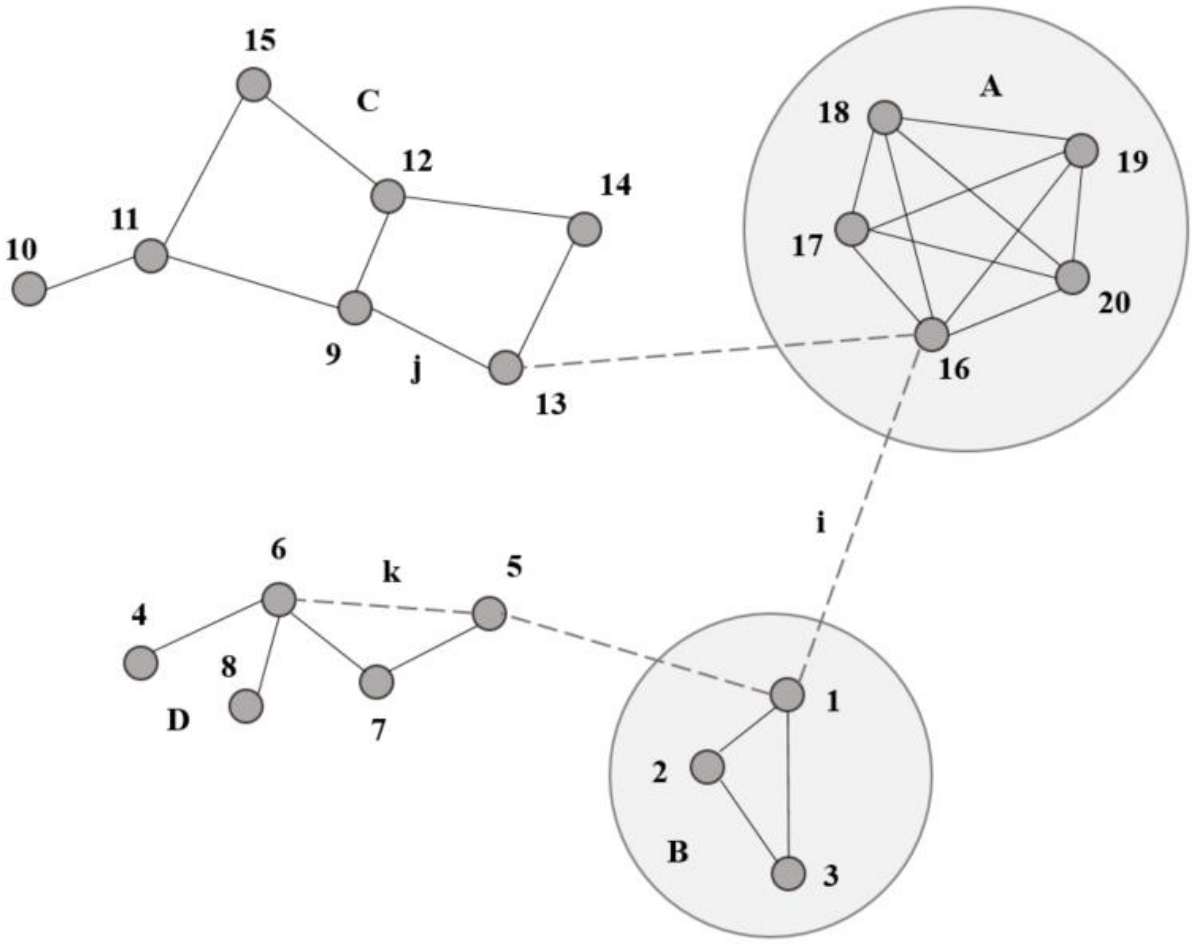

Рисунок 1.1 – Пример малой сети распространения информации

Грановеттер [1] предложил классификацию ссылок, основанную на силе представленных отношений. Классификация включает в себя слабые связи («знакомство»), прочные связи («дружба») и мосты, где мост - это канал в сети, который обеспечивает единственный путь между двумя произвольными узлами (канал I на рисунке 1 является мостом). Надежность канала связи определяется множеством факторов, таких как постоянство, взаимность, частота и продолжительность сообщений, а также продолжительность существования. Сильные связи в массовых онлайн сетях указывают на наличие сильных онлайн-отношений между участниками сети, в то время как слабые связи вовсе не обязательно соответствуют каким-либо онлайн-отношениям. Было замечено, что если два узла (1 и 2 на рисунке 1.1) сильно связаны с другим узлом, то, с очень высокой вероятностью, они также сильно связаны друг с другом: «друг моего друга - мой друг». Этот механизм называется триадическим замыканием. Более того: чем сильнее связь между двумя членами сети, тем больше степень совпадения их дружеских кругов и тем больше вероятность закрытия.

Необходимо обратить внимание, что триадическое замыкание не работает для слабых связей («знакомый моего знакомого не обязательно является моим знакомым»). Из-за триадного замыкания группа узлов, которые

имеют некоторые сильные связи, имеет тенденцию постепенно формировать все более сильные связи и в конечном итоге становится полностью и сильно связанным кластером (группы A и B на рисунке 1.1). В кластере каждый является «другом» каждого. Как показал Льюис [11], эта плотная социальная близость делает кластер идеальным общим источником знаний: частые контакты между членами гарантируют, что кластер быстро узнает новый факт. Если этот факт забыт или искажен, добровольно или невольно, членом группы, он может быть быстро и надежно восстановлен.

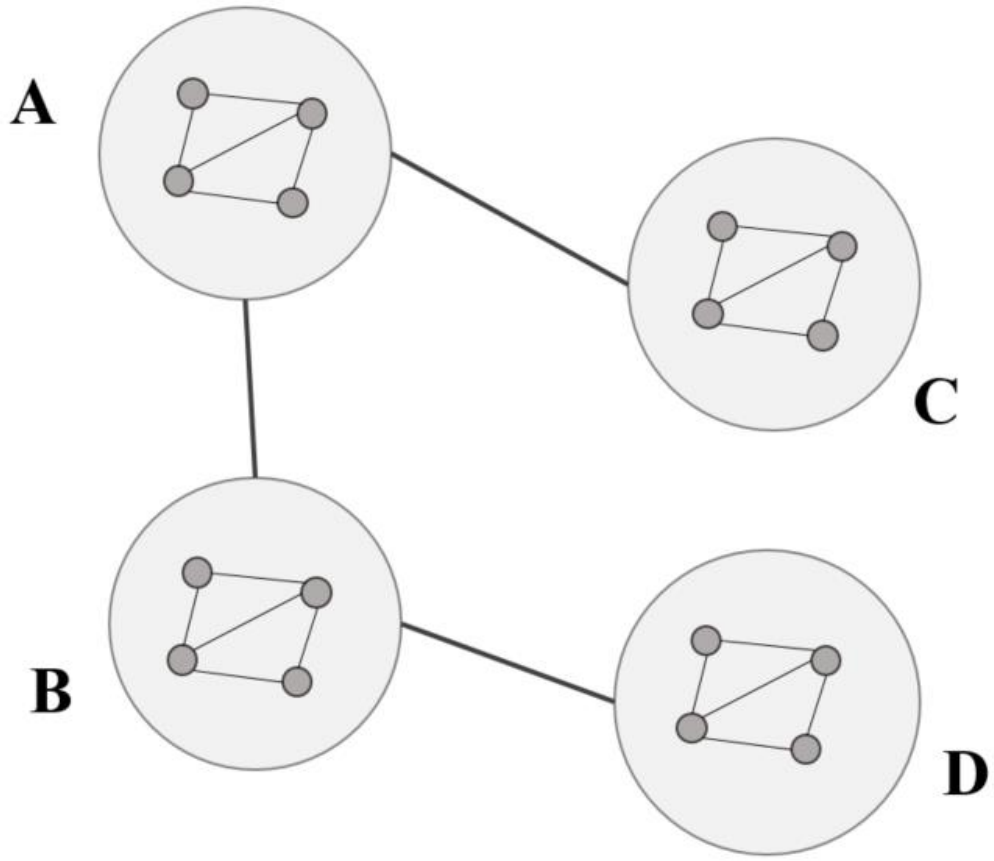

Рисунок 1.2 – Сильные связи между кластерами, а слабые внутри них

Ограничением сильных связей является их короткий диапазон распространения: если вся социальная сеть не является единым кластером, неизбежно будут регионы, не затронутые распространением. Относительное число кластеров в сети характеризуется коэффициентом кластеризации C - средним числом триадных замыканий в непосредственной близости от рассматриваемого узла, которые не включают сам этот узел. C - это число в диапазоне от 0 до 1: C = 1 для сети, которая формирует один связанный между собой кластер, и C = 0 для дерева (сеть без связей между ветвями). Для большинства реальных социальных сетей C меньше 0,5, что означает, что они сформированы более чем одним кластером, и сильные связи не действуют как эффективные каналы.

Слабые связи, которые перекрывают «разрывы» между кластерами, необходимы для распространения информации в реальных сетях. Слабые связи создают дополнительные более короткие пути между пулами знаний и позволяют распространять информацию, с искажениями или без них. Слабые связи расширяют возможности соседних с ними членов сети: люди с множеством слабых связей лучше всего могут использовать информацию. Это, в свою очередь, улучшает социальный статус этих людей. Подводя итог, можно сказать, что в социальной сети существует разделение труда (рисунок 1.2): сильные связи ответственны за формирование кластера, которые, в свою очередь, служат общими пулами знаний, а также источниками доверия и влияния («мозги»); слабые связи действуют как информационные и инновационные каналы («нервы»).

1.1.2 Типы распространяемых сообщений

В процессе распространения информации необходимо уделить отдельное внимание тому, что является объектом передачи, а именно: что представляет из себя сообщение. При рассмотрении контента сообщения, необходимо отметить, что каким бы ни был контент, он, как правило, подвержен определенным искажениям на пути распространения, как показано в исследовании Харари [12]. Тем не менее, важно уделить внимание структуре передаваемого сообщения.

Доминирующим способом распространения информации в офлайновых сетях является аудиосвязь: лицом к лицу, по двусторонней радиосвязи или по телефону. Письменные сообщения (рукописные и печатные буквы, телеграммы и факсимильные сообщения) все еще используются, но быстро заменяются электронными сообщениями. Несмотря на то, что они отличаются с физической точки зрения, все аудио- и письменные сообщения в основном несут текстовую информацию (возможно, с некоторыми графическими изображениями) и предназначены для одного получателя (в конференц-звонках участвуют несколько разговаривающих людей, но размер аудитории редко превышает десять человек). Другими словами, общение обычно бывает только текстовым, пиринговым и частным.

Напротив, онлайновые сети используют различные коммуникационные механизмы, как напрямую, так и через веб-сайты социальных сетей. Прямая связь включает электронную почту (с вложениями мультимедиа или без них; адресованные одному получателю или списку рассылки) и мгновенные сообщения (Skype, Google Talk, Microsoft Teams и т.д.). Непрямые коммуникации осуществляются через встроенные механизмы обмена сообщениями в социальных сетях (личные сообщения), а также сообщения различных сообществ и СМИ.

Некоторые сайты социальных сетей позволяют членам сети в определенной степени контролировать получателей сообщений групповой рассылки (например, «только друзья», «друзья и друзья друзей», и «Все» в

Facebook). Групповая рассылка и, особенно, широковещательные сообщения повышают скорость распространения - при условии, конечно, что они не игнорируются предполагаемыми получателями. Интересный класс сообщений, которые, кажется, существуют исключительно в массовых онлайн-социальных сетях через сайты социальных сетей, - это пинговые сообщения, также известные как бессодержательные сообщения.

Пинговые сообщения были впервые изучены Кайе Дж. [4] в отношении компьютерного приложения Virtual Intimate Object, которое позволяет двум людям оставаться в прямом контакте онлайн, нажав кнопку на панели инструментов на рабочем столе компьютера. Нажатие кнопки заставляет аналогичную кнопку на удаленном компьютере мигать или менять цвет, что позволяет другому человеку узнать, какую информацию ему хотят передать. Несмотря на распространенное мнение, выраженное, например, Голдером. [6], пинговые сообщения не всегда несут только один бит информации. Количество информации в пакете составляет $I = 1 - \log_2(1 - L)$, где L - коэффициент потерь сквозного канала связи между вовлеченными компьютерами, то есть доля всех отправленных сообщений, которые не имеют статуса «Получено». $I = 1$, только если соединение осуществляется без потерь ($L = 0$). Единственное важное свойство всех сообщений, упомянутых выше, заключается в том, могут ли они быть широковещательными или, по крайней мере, многоадресными. Далее предполагается что сообщения имеют только одного получателя, если не указано другое.

1.2.2 Восприятие информации агентами распространения

Сообщение с новой информацией, полученное человеком в сети распространения информации, не обязательно приводит к немедленному принятию этой информации человеком. По крайней мере, два класса моделей объясняют механизмы принятия: поэтапное распространение, основанное на коллективных действиях, изученных Чве [13], и диффузия, основанная на моделях заражения, предложенных Кермаком и МакКендриком [14] и другими авторами. Люди в постановочных моделях распространяют и чувствуют готовность других принять новые знания. Их девиз: «Я пойду [за чем-то], если ты пойдешь [за этим]». Если формируется критическая масса получателей, они вместе принимают знания и, возможно, становятся новыми источниками информации. В противном случае новые знания коллективно отвергаются.

Чтобы сформировать критическую массу, людям важно иметь достаточную информацию о намерениях, по крайней мере, некоторых других участников сети: они планируют «восстать» или «принять» новые знания? Естественно, эта информация может быть собрана участником только путем обмена сообщениями с непосредственными соседями. Обратите внимание, что соседи находятся в аналогичной позиции: прежде чем они «восстанут» или «примут», они также лучше узнают намерения своих соседей (или используют свою интуицию, которая может быть смоделирована с использованием

математической теории игр). У этой рекурсивной проблемы есть по крайней мере одно хорошее решение: кластер, который уже упоминался выше. Действительно, в кластере все члены имеют непосредственный прямой доступ ко всем другим членам и находятся в лучшем положении, чтобы узнать их намерения - таким образом формируя общее мета-знание о том, принимать или не принимать новые знания.

Общие (мета) знания являются результатом постепенного и разнородного процесса. По словам Чве [13], он начинается с первоначальных принимающих информация: «ненормальных» членов социальной сети, которые намерены принять новые знания. Первоначальные принимающие, не являются специальными, преданными членами сети: они просто взяли на себя эту роль в этом конкретном процессе распространения, даже если при этом одни и те же люди могут быть психологически предрасположены к тому, чтобы инициировать распространение информации снова и снова. Первоначальные принимающие обычно образуют собственный кластер - «ведущий». Информация распространяется вдоль слабых связей от ведущего кластера через цепочку других кластеров: от первоначальных новаторов до последователей, а затем от последователей до поздних последователей (Рисунок 1.3). Постановочная модель еще раз показывает, что слабые связи более важны для распространения информации, в то время как сильные связи поддерживают доверие, знания и построение коллективных действий. Постановочная модель дает некоторое представление о динамике распространения информации. Тем не менее, она не учитывает забывчивость членов социальной сети: согласно Чве [13], когда факт узнается сообществом или отдельным лицом, он никогда не забывается и не может быть повторно изучен. Инфекционные модели диффузии, частично рассмотренные Китцаком М. [15] рассматривают это явление подробнее.

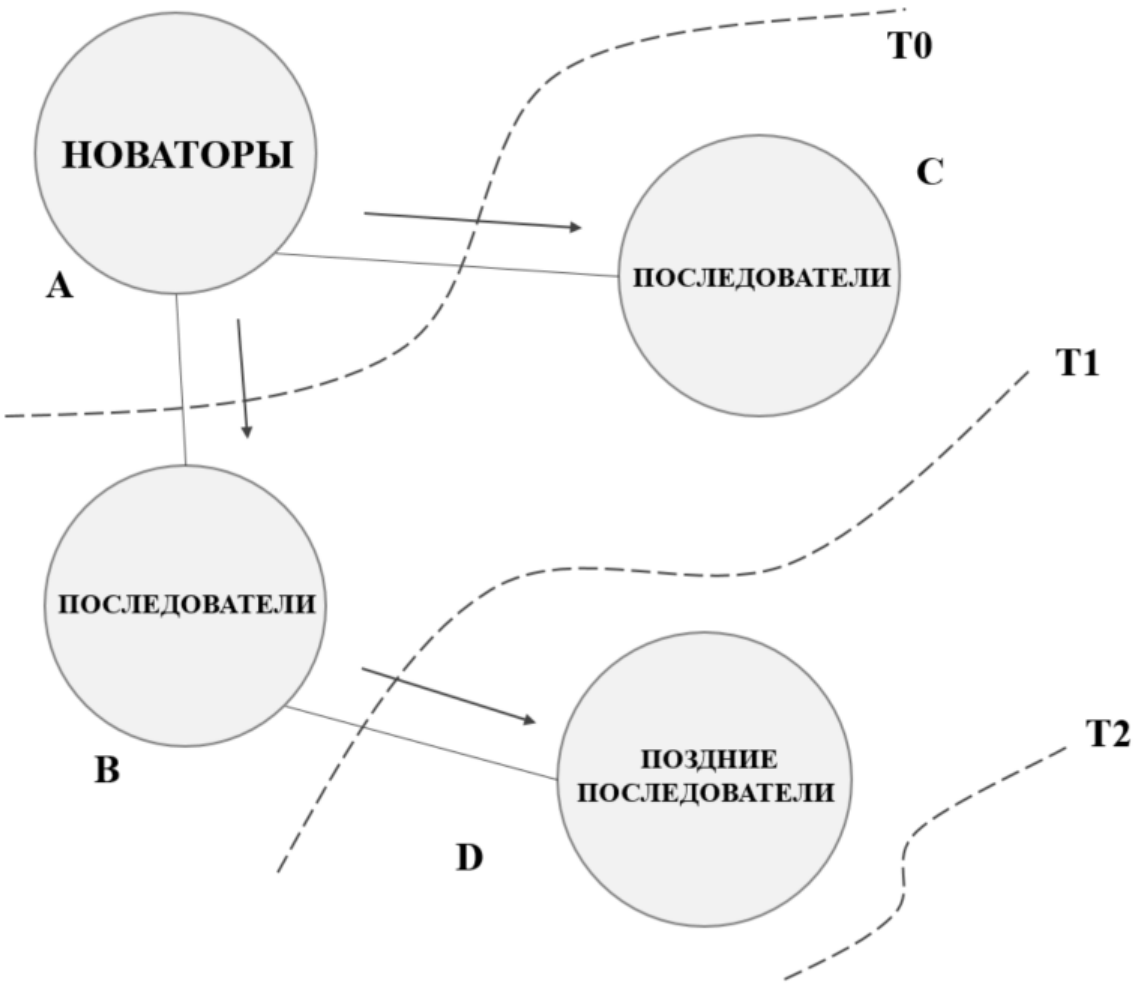

Рисунок 1.3 - Модель распространения информации в распределенной сети.

Существуют две основные инфекционные (или сплетничающие) модели, относящиеся к Кермаку и МакКендрику [16]: восприимчиво-инфекционно-выздоровевшие (SIR) и восприимчивые-инфекционно-восприимчивые (SIS). В моделях используется медицинская или, скорее, эпидемиологическая) терминология и метафоры для описания распространения информации. Каждая модель на самом деле не является моделью, а представляет собой группу похожих моделей, которые описывают мелкие детали. В модели SIR каждый узел в социальной сети может находиться в одном из следующих состояний: восприимчивый (S), зараженный (I) или выздоровевший (R). Первоначально все узлы восприимчивы - соответствующие люди не знают факта, который распространяется, и поэтому могут стать «зараженными», изучая его. Необходимо обратить внимание, что, в отличие от эпидемиологии, в моделях SIR или SIS нет узлов, невосприимчивых к информации. Это ограничение можно легко устранить, добавив другое состояние - иммунное. На каждом временном шаге инфекционные узлы пытаются заразить восприимчивые узлы в их окрестности сети с некоторой вероятностью и затем входят в выздоровевшее состояние. Заражение узла означает передачу новой

информации этому узлу. Выздоровление означает становление иммунного постфактума.

Восстановленные узлы не остаются заразными - они не могут далее распространять информации. В модели SIS после заражения узел на каждом временном шаге либо остается заразным с вероятностью 1 -λ, либо снова становится восприимчивым с вероятностью λ. Последняя ситуация соответствует забвению информации: после неявного восстановления восприимчивый узел может быть снова «заражен» той же информацией. Богуна и его соавторы [7] доказали, что модель SIS имеет существенный недостаток: у нее нулевой эпидемический порог. Это означает, что факт, введенный в сеть, описанную с использованием модели SIS, в конечном итоге становится известным каждому, что явно ложно во всех реальных социальных сетях. Ни инфекционные, ни поэтапные модели полностью не описывают процесс принятия информации в социальных сетях. Новая модель, свободная от их ограничений, еще не предложена.

1.1.3 Определение влиятельных членов распространения информации

Эффективное распространение в социальных сетях было бы невозможно, если бы влиятельные распространители (то есть «новаторы» или «ранние последователи» новых знаний) не находились внутри сети. Например, узел 4 на Рисунке 1.1 вряд ли может быть существенным распределителем: все сообщения из 4 узла должны быть получены и затем повторно переданы соседям узлом 5, при этом если индивид, представленный узлом 5, не отвечает, занят или просто находится в плохом настроении, то распространение прекратится до того, как оно даже может начаться. Возникает вопрос: как возможно определить, предпочтительно, просто посмотрев на график социальной сети, что узел 4 плохой распространитель? Кроме того, возможно ли определить истинные в распространителях, просто посмотрев на сетевой граф? Чве [13] заметил, что, по крайней мере, влиятельные распределители не должны быть слишком тесно связаны и должны быть оптимально распределены: если распределители разнесены по сети слишком тонко, это считается что они излишне «распылены», если же наоборот сконцентрированы, то это можно назвать что они «геттоизированы». Как уже видно из моделей, связь по длинным путям невозможна, потому что информация может быть искажена, и существует растущая вероятность того, что сообщение будет отброшено «иммунным», не взаимодействующим узлом.

Узел распространителя действует как фонарь, он словно включенный «информационный прожектор» в окрестности своей сети. И если лучи «прожекторов», распространяемые разными участниками, не пересекаются, глобальное распространение невозможно. Это ограничение приводит к «атомизации». Причина «геттоизации» очень похожа: если два распространителя расположены вместе, их «прожекторы» почти полностью

перекрываются, оставляя остальную часть сети «неосвещенной». Другими словами, топология социальной сети действительно играет роль: один узел на периферии имеет минимальное влияние, как и плотно упакованные или слабо рассредоточенные узлы. Для количественной оценки относительной важности личности в асоциальной сети была предложена концепция центральности.

Центральность - это число, обычно от 0 до 1, показывающее, в какой степени узел, представляющий индивида, является «центральным». Используются различные типы центральности: степень центральности (число соседей узла), центральность близости (среднее расстояние от узла эго до всех других узлов в сети), центральность между промежутками (доля всех кратчайших путей в сети, проходящих через узел) и т. д. Все факторы центральности могут быть формально рассчитаны по сетевому графу, хотя вычисления могут быть очень сложными, в зависимости от размера сети.

По мнению Фримена [16] и других исследователей, именно высокая центральность делает узел эффективным распространителем. Действительно, высокая степень центральности означает, что у человека много соседей, и он может сразу сделать информацию доступной для всего своего круга друзей. Центральность высокой степени близости подразумевает, что пути от индивидуума к остальной части сети короткие, что делает процесс распространения беспрепятственным, подверженным искажениям и затуханию. Наконец, считается, что высокий уровень межличностных отношений определяет, кто имеет больше «межличностного влияния» на других.

Однако Китцак [15] предполагает, что критерии центральности эффективности действительны лишь в некоторой степени, и предлагают другой количественный подход к проблеме идентификации влиятельных распределителей, основанный на k-ядрах. K-ядро (или k-оболочка) сети - это набор вершин K, где каждая вершина в K имеет как минимум k связей с другими вершинами в K (Рисунок 1.4). Число k называется индексом ядра. Ядра с более высокими индексами являются подмножествами ядер с более низкими индексами. Размер k-ядра резко уменьшается при больших значениях k. Китцак [15] показывают, что ни степень, ни центральность между ними не играют важной роли в прогнозировании эффективности узла в качестве распределителя. В случае одного распространителя на размер «зараженной» популяции в основном влияет k_S - индекс наименьшего ядра, содержащего распространяющие узлы. Узлы с наибольшим значением k_S, вероятно, будут лучшими распространителями. Коррелирует ли k_S с центральностью близости, пока неизвестно.

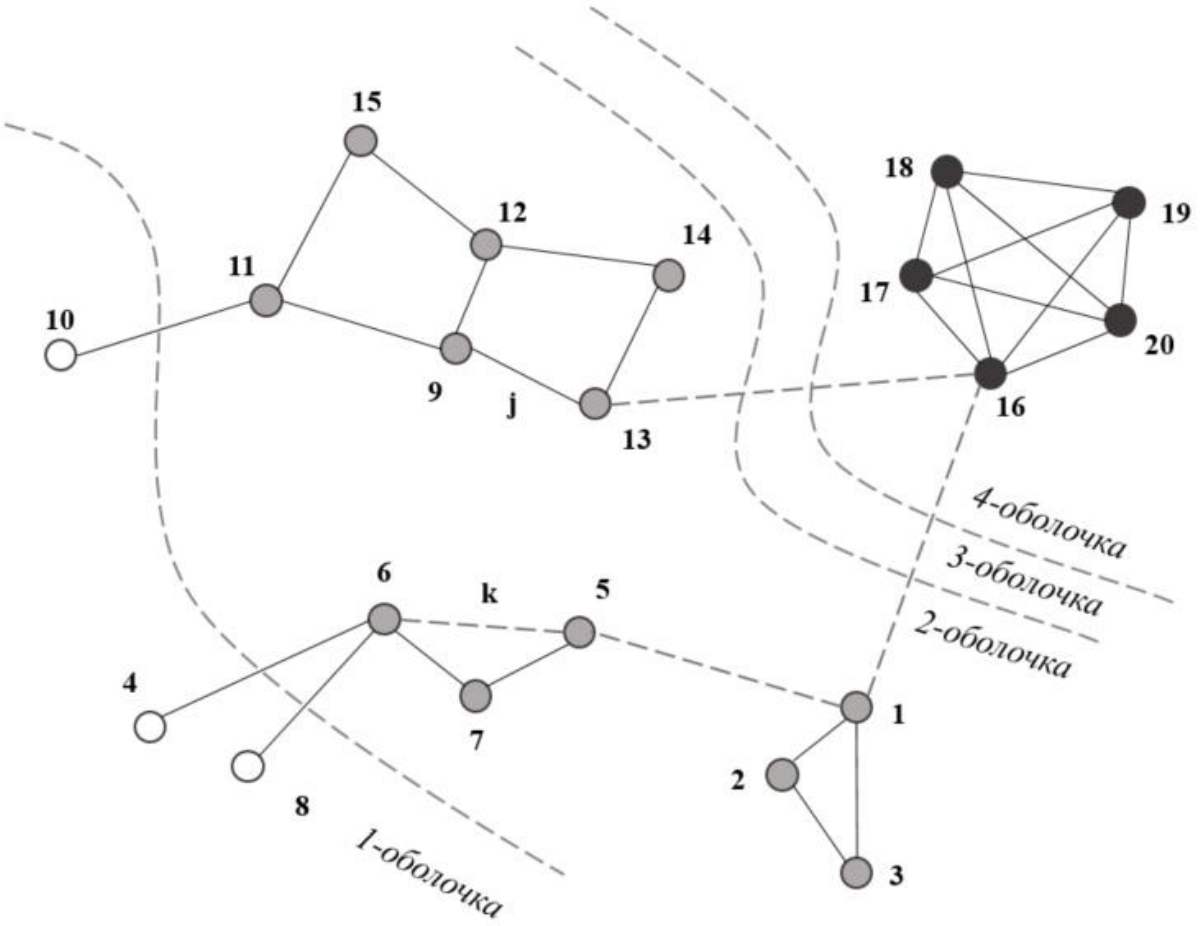

Рисунок 1.4 - 1-оболочка, 2-оболочка, 3-оболочка и 4-оболочка в социальной сети. 3-оболочка и 4-оболочка совпадают.

Ситуация меняется, если имеется более одного распределителя, причем все распространители взаимодействуют с одной и той же информацией коррелированным (синхронным) способом. В этом случае лучшей стратегией распространения является выбор либо узлов с наивысшей степенью, либо узлов с наивысшим индексом оболочки k_S с требованием, чтобы никакие два из них не были напрямую связаны друг с другом (чтобы избежать «геттоизации»). интересная связь между описанной выше моделью сплетен SIS и k-оболочками: если в сети SIS есть самовоспроизводящийся информационный «вирус» такой как слух, городской миф или история ужасов, затем, согласно Китцаку, он сохраняется в основном в слоях с высоким индексом. Практическое следствие этого наблюдения заключается в том, что для подавления распространения нежелательной информации необходимо уделять особое внимание внутренним оболочкам сети.

1.1.4 Динамика распространения сообщений

В модели, которая представляет сеть в виде графа, предполагается, что все узлы и все ссылки равны: в частности, каналы связи, представленные

ссылкой, имеют одинаковые (или похожие) свойства и имеют значение только топологические свойства графа в целом (например, меры центральности и длина пути. Это явно не так: по крайней мере, мы знаем, что ссылки могут быть сильными и слабыми, и это различие, хотя и существенное с точки зрения распространения, обычно не отражается в сетевом графе.

При близком рассмотрении связь между двумя гипотетическими членами сети, А и Б, является сложным объектом, который имеет некоторые собственные свойства (физическая реализация, скорость или задержка передачи, максимальная емкость канала, надежность или вероятность потери сообщения, величина искажения данных и т. д.) и неявно разделяет некоторые свойства А и Б. Мы уже видели, что любой участник может находиться в одном из режимов принятия информации, в зависимости от модели распространения. Например, Б может быть «восстановлен» из факта F, то есть он не принимает никаких новых сообщений, связанных с F; в этом случае ссылка, соединяющая А и Б, несет только выборочную информацию, а именно: все, кроме F.

Более важным свойством реальных связей в распределенных сетях является то, что они не являются непрерывными. Косинец [8] отмечает, что информация распространяется только в результате дискретных коммуникационных событий, таких как электронная почта или текстовые сообщения, разговоры, встречи или телефонные звонки, которые распределяются неравномерно во времени. Именно неравномерное распределение сообщений делает распространение практически непредсказуемым: просто невозможно заранее сказать, когда та или иная связь будет использоваться для передачи, и даже будет ли использоваться вообще.

Из неоднородности распределения сообщений есть следствие: если ссылка интенсивно используется для связи, то это, вероятно, сильная связь. Поэтому вряд ли она будет находиться на активном пути распространения информации. Обратное утверждение верно относительно слабо используемой (слабой) связи. Другими словами, больше сообщений - меньше информации, и наоборот.

Груль [10] и Голдер [6] изучали динамику сообщений в агрегированных интернет-блогах и Facebook - крупнейшей на сегодняшний день сети распространения информации. Они обнаружили, что обмен сообщениями в сети, хотя и подвержен неоднородным шаблонам, не является полностью случайным. Он демонстрирует устойчивые и последовательные временные ритмы - по крайней мере, в университетских городках, используемых для анализа, и во время определенных сезонов. Тысячи индивидуальных решений, сделанных членами сети, вместе выглядят так, как будто они поставлены невидимым режиссером. Интересно, что синхронность имеет место не только при изучении «классических» сообщений с текстовыми и, возможно, мультимедийными вложениями, но и бессодержательными сообщениями – пингами. Удалось выяснить, что они следуют тем же временным шаблонам.

Частоты сообщений (и объемы) в сети следуют за тремя основными циклами: ежедневный цикл, недельный цикл и годовой (сезонный) цикл. Это

неудивительно, так как обмен сообщениями является неким прокси для онлайн-активности в социальных сетях, и если члены сети ведут схожий образ жизни их шаблоны общения также будут схожими.

Цикл ежедневной активности в социальных сетях имеет минимум в 3–8 утра, за которым следует полуденный подъем в полдень и «плато» между полуднем и полуночью. Стоит сравнить это с ежедневным циклом обмена сообщениями в корпоративной сети электронной почты, а не в социальной сети: существуют два максимума в 11:00 и 17:00 и минимум с 19:00 до 07:00. Как социальная сеть, так и корпоративная сеть продемонстрировали высокую активность членов в течение первых нескольких дней недели и минимальной активности в выходные дни. Наконец, люди в обоих типах сетей были гораздо более активными в июне-августе и декабре-январе, чем в оставшиеся сезоны.

В дополнение к периодическим колебаниям, частоты сообщений подвержены временному экспоненциальному затуханию: из-за социальной апатии уровень активности пар пользователей имеет тенденцию заметно снижаться с течением времени (например, до 70% активных связей в Facebook умирают в течение одного месяца). Только небольшая часть ссылок сохраняется в течение значительного времени.

Крайним случаем периодических колебаний частоты сообщений являются бесконечно длинные циклы, которые имеют место в относительно плохо изученных одноразовых сетях. В одноразовых сетях обмены происходят один раз между людьми, которые не ожидают повторного обмена в будущем. Примерами одноразовых сетей являются электронные площадки типа Ebay - веб-сайт онлайн-аукциона. В одноразовых сетях неизвестно, будут ли два человека подключены в будущем, и если да, то когда. Аналогичным образом, наличие связи между двумя узлами в одноразовом сетевом графе не имеет большого значения: это указывает на то, что в прошлом было взаимодействие между узлами, но неясно, будут ли узлы когда-либо взаимодействовать в будущем.

Изучение одноразовых сетей включает использование математической теории игр, которая количественно определяет намерения и ожидаемые выгоды отдельных лиц от коммуникации и пытается предсказать, возможно ли взаимодействие между двумя членами сети в будущем. Эмануэльсон и Уиллер [17] предложили, что одноразовые и повторяющиеся обменные структуры различны, и их диффузионные структуры также различны. Однако на данный момент нет систематической теории, объясняющей диффузию в одноразовых сетях.

Относительная непредсказуемость частот связи приводит к непредсказуемости сквозной задержки сообщения. Поскольку средняя длина кратчайшего пути в типичной социальной сети близка к шести ссылкам (феномен «шести степеней разделения»), средняя задержка связи между парами узлов может достигать восьми дней. Эта задержка замедляет распространение информации и сильно влияет на конфигурацию предпочтительных каналов связи.

1.1.6 Гомофильность в сетях распределения

Гомофилия - это склонность индивидов к общению с другими людьми. Структура путей распространения в социальных сетях сильно зависит от гомофильности среди пользователей: люди больше общаются и дольше общаются с людьми, которые похожи на них. Другими словами, если в графе социальной сети есть связь, представляющая связь между двумя людьми, есть веская причина полагать, что они имеют что-то общее между собой. Кроме того, чем больше у них общих черт, тем сильнее связь между ними со всеми вытекающими последствиями для распространения информации.

Гомофильность в современных социальных сетях изучали Голдер [6] и Лесковетц и Хоровитц [18]. Вместо наблюдения за отдельными сообщениями, которые могут быть случайными и часто запускаемыми механизмами, самой сети, можно группировать последовательные взаимные сообщения в сеансы. Разговор с несколькими сообщениями является лучшим показателем гомофилии, а продолжительность сеанса (по количеству сообщений или по времени) может использоваться в качестве количественной меры гомофилии.

Самая сильная гомофилия среди пользователей крупных социальных сетей существует в отношении используемого языка, совместного посещения школы (в Facebook), географического местоположения разговоров, а затем возраста, но не пола: люди чаще общаются чаще и с большей продолжительностью общения с противоположным полом. С возрастной точки зрения молодые люди (в возрастной группе от 14 до 35 лет) в основном общаются с людьми примерно того же возраста, как примечание, они также говорят быстрее, в то время как более старые члены сети разговаривают дольше и отправляют больше сообщений за сеанс.

Гомофилия, основанная на географических местоположениях (или, скорее, на расстоянии между местоположениями связанных людей), следует четкому взаимному закону: количество разговоров уменьшается с расстоянием примерно в 1/x. Продолжительность разговора также уменьшается с расстоянием, однако количество обмениваемых сообщений остается постоянным до медленного уменьшения. В совокупности эти наблюдения подразумевают, что большая часть трафика в современных сетях распространения информации сконцентрирована в географически компактных областях - возможно, в сильных сообществах, с большим коэффициентом связности.

1.1.7 Маршрутизация сообщений

С практической точки зрения одной из наиболее важных проблем, связанных с распространением в социальных (и других) сетях, является маршрутизация сообщений: как рассчитать наиболее эффективный путь (маршрут) от источника сообщения к произвольному месту назначения? (Или ко всем сетевым адресатам одновременно.) Во многих реальных сетях

проблема обычно решается путем введения иерархической схемы адресации или метрической системы координат.

Первый подход используется, например, в Интернете. Четырехбайтовый (IP v4) или 16-байтовый (IP v6) IP-адрес - это уникальный идентификатор, назначенный сетевому объекту; это комбинация сетевого адреса и адреса хоста объекта. Хост - это обычно отдельный компьютер, а «сеть» - это совокупность хостов, напрямую или тесно связанных друг с другом. В терминологии сети распространения информации хост - это узел, представляющий человека, а сеть - это нечто вроде кластера. Навигация (маршрутизация) происходит следующим образом: сначала необходимо обнаружить сеть назначения (используя так называемые таблицы маршрутизации), а затем связаться с хостом в этой сети.

Сайты социальных сетей, например, такие как Facebook и Twitter, назначают числовые идентификаторы сетевым узлам. Однако эти идентификаторы являются просто последовательными или случайными числами. Они не связаны ни с положением узлов в сети, ни с их демографией и не могут использоваться для навигации

Метрическая система координат систематически связывает набор координат с каждым узлом в сети. Хорошо известные методы оптимизации (такие как градиентный спуск) могут затем использоваться для нахождения кратчайшего пути через сеть от источника к месту назначения. Примером эффективно реализованного метрического пространства в сети является схема именования улиц и проспектов в манхэттенском стиле.

Гипотетическая метрическая система координат отделена от скрытого метрического пространства, невидимого для большинства членов сети и даже для внешних наблюдателей, но которое является существенным для принятия решений о маршрутизации. То, что такое скрытое метрическое пространство может существовать, неявно следует из известного эксперимента Милграма о маленьком мире [19], где случайных людей просили отправить по почте открытку неизвестному адресату, отправив ее одному из своих ближайших друзей или знакомых. Очевидно, что участники использовали некоторое руководство (которое мы можем назвать либо интуицией, либо смыслом скрытого метрического пространства), чтобы решить, какой из их ближайших партнеров по социальной сети является наиболее подходящим следующим пунктом назначения.

Ожидается, что скрытое метрическое пространство будет иметь хотя бы одно свойство: чем меньше расстояние между любыми двумя узлами в скрытом пространстве, тем более вероятно, что они связаны в наблюдаемой (реальной) топологии графа сети. Если гомофилия действительно существует, то вероятность соединения двух узлов зависит от сходства между узлами, которые, в свою очередь, зависят от демографии и других внутренних психологических характеристик узлов.

В настоящее время не существует систематического способа построения скрытого метрического пространства для произвольной массовой онлайновой

сети. Выводы из работы Богуны[7] заключаются в том, что хорошая навигация в целом зависит от четкого безмасштабного (степенного) распределения степеней узлов и высокой кластеризации. Если расстояние между двумя произвольными узлами определено как (kk ′) α, где k и k ′ - степени обоих узлов, а α - коэффициент кластеризации, то схема маршрутизации с уменьшением / уменьшением (ZOZI) может быть использована следующим образом: сообщение сначала отправляется от исходного узла A на некоторый узел высокой степени («концентратор») в непосредственной близости от A, затем от этого «концентратора» на другой «концентратор» в окрестности узла назначения B и оттуда до B. Для работы принципа ZOZI сеть должна иметь достаточно много «хабов», что обычно не имеет место в реальных сетях распространения.

1.1.8 Актуальные информационные пути

В общем и целом, ссылки в графе сети не являются достоверными индикаторами реального взаимодействия между узлами. Они часто отображаются для представления статуса и идентичности членов сети, представленных соседними узлами. Они не обязательно показывают уровень взаимного доверия, общих интересов или чего-либо общего - даже информацию (например, Уилсон [20] пишет, что за все время наблюдения только 50% связей на Facebook использовались для взаимодействия).

Таким образом, ссылки на сетевой график могут использоваться для прогнозирования распространения информации только с оговорками. Лучшей альтернативой графу сети является граф взаимодействия (также известный как сеть активности). Это граф G ’(V, E’), который состоит из вершин V и ребер E ’. Граф активности имеет те же узлы, что и исходный граф социальной сети; а социальная связь между двумя узлами A и B существует в G ’тогда и только тогда, когда пользователи, представленные узлами, взаимодействуют напрямую через общение или приложение. При этом необязательно, чтобы соответствующая связь между A и B существовала в исходном сетевом графе. Некоторые исследователи предполагают, что для построения графа взаимодействия рассматриваются только самые последние взаимодействия.

График взаимодействия отражает действительные взаимодействия между людьми. Его ссылки являются реальными каналами связи (в отличие от ссылок на сетевом графике, которые являются потенциальными каналами, потенциал которых может быть так и не быть реализован)

Поскольку многие ссылки в реальном сетевом графе используются слабо или никогда не используются, соответствующий граф взаимодействий обычно имеет меньше ссылок. Это означает, что степень взаимодействия D_I - количество каналов взаимодействия на узел - обычно ниже, чем социальная степень D_S того же узла. Уилсон [20] графически представляет экспериментальную зависимость между двумя степенями, из которой следует, что $D_I \sim \pi D_S$, где $\pi \ll 1$. Кроме того, график взаимодействия демонстрирует более

точное степенное масштабирование, имеет большую длину пути, меньше «суперузлов» («концентраторов») и меньшую кластеризацию. Последние два свойства - по крайней мере теоретически - понижают эффективную маршрутизацию в частности и распространение в целом.

Косинец [8] продемонстрировал, что структура графа взаимодействия не случайна: граф является самодостаточной структурой, которая построена вокруг высокоэффективных связей в исходном сетевом графе и способствует высокой эффективности связей.

При рассмотрении сети на рисунке 1.1 еще раз можно увидеть, что если узлы 5 и 6 не взаимодействуют активно, а узлы 6 и 7 и 5 и 7 активны в паре, то связь между двумя узлами относительно неэффективна: для того, чтобы 5 узнал новости непосредственно от 6, потребуется больше времени, чем косвенно с 6 и с 7 (кто, в свою очередь, получил бы ее с 6). Информационные пути в реальных социальных сетях с односторонними ссылками не всегда прямые. Кроме того, сообщение, которое принимает прямой путь, может быть устаревшим. Это явление нарушает неравенство треугольника, и в структуре общения в реальных социальных сетях преобладают такие нарушения.

Чтобы различать график взаимодействия, который показывает какие взаимодействия имели место в недавнем прошлом, от фактических и текущих путей сообщения, последние называются «магистралью сети». Магистраль сети в момент времени T - это подмножество связей в графе взаимодействий, которые не проходят мимо более быстрого альтернативного пути - другими словами, связей, которые не нарушают неравенство треугольника. Практически магистраль представляет собой очень разреженный подграф, состоящий как из глубоко встроенных ссылок, так и из мостов большой дальности.

Ясно, что из-за динамического характера шаблонов сообщений в социальной сети магистраль не всегда будет одинаковой: некоторые ссылки могут замедляться «на обед», а другие могут ускоряться «на деловую встречу». В социальной сети, охватывающей несколько часовых поясов, этот сценарий не только возможен, но и весьма вероятен. Магистраль сети в момент времени T + ΔT может сильно отличаться от магистрали в момент времени T. Тем не менее, как ни удивительно, каждая мгновенная магистраль использует в среднем 75% ссылок совокупной магистрали - объединение всех мгновенных магистралей за некоторое время.

Прогнозирование структуры в сетях информации имеет решающее значение для эффективного распространения. Расположение влиятельных участников зависит от топологии сети, но внесение динамики связей в картину может кардинально изменить их влияние. Можно оценить насколько изменится влияние «суперпользователя», который имеет тысячи соединений и переходит в самую внутреннюю k-оболочку, но ссылки используются только раз в месяц.

1.1.9 Секретность и безопасность информации

Анализ распространения информации в распределенных сетях будет неполным, если не рассмотреть вопросы секретности.

Хотя верно и то, что самые массовые онлайн-сети были разработаны с учетом секретности, при этом некоторые реальные социальные сети (как онлайн, так и другие) чрезвычайно чувствительны к этому фактору. В сетях, где имеется необходимость в секретности, важно поддерживать баланс между производительностью информации (скорость распространения и охват) и секретностью как таковой. Для количественной оценки секретности можно оценить вероятность обнаружения канала и вероятность экспозиции узла (которые, в свою очередь, являются монотонными функциями от количества каналов: чем больше ссылок, тем выше вероятность обнаружения и вероятность воздействия). К сожалению, меньшее количество ссылок означает меньшее количество путей распространения, что ухудшает распространяемость информации. Проблема секретности может иметь оптимальное решение, которое улучшает безопасность, но поддерживает разумную производительность и наоборот, но в целом полностью сбалансированного решения не существует.

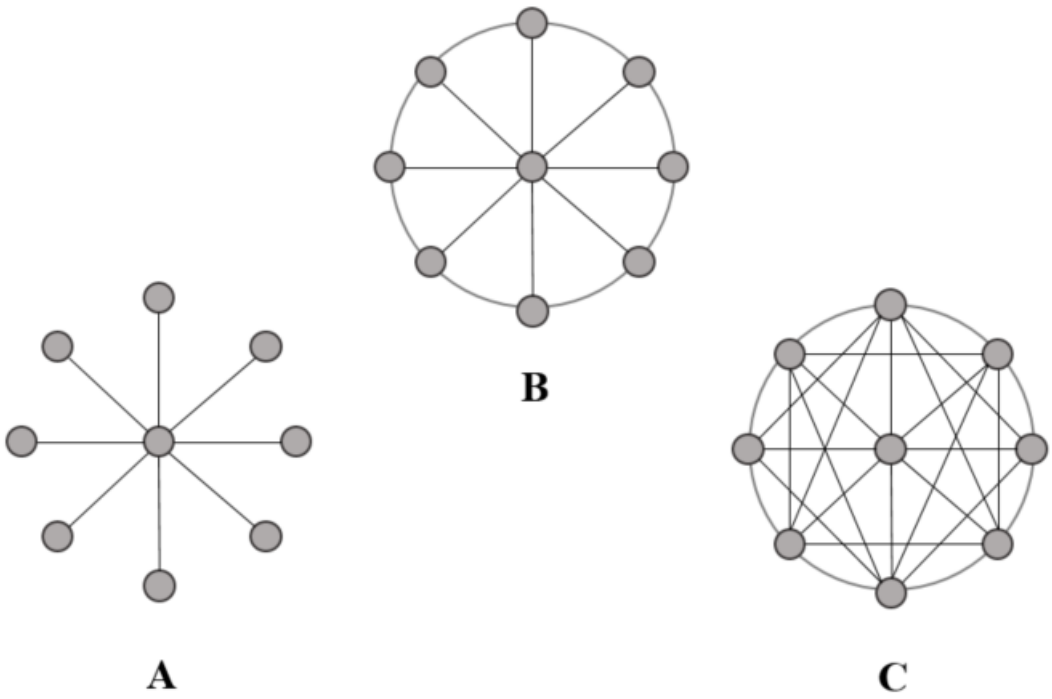

Рисунок 1.5 - Оптимальные структуры защищенных сетей: (A) топология звезда, (B) усиленное кольцо и (C) полный граф

Секретность в социальных сетях - плохо изученная тема. Одно из немногих исследований в этой сфере это исследование Линде [21]. Авторы применили математическую теорию игр, чтобы найти сильную зависимость

оптимального решения от топологии сети (не рассматривая динамику связи). Другим решающим фактором является этап скрытых операций, которые выполняются сетью.

На начальном этапе, когда сетевая иерархия еще не зафиксирована, оптимальная структура связи соответствует либо графа в виде звезды (сеть с одним центральным «жирным узлом»), либо полного графа (универсальная сеть), в зависимости от вероятности обнаружения канала: звезда подходит для менее безопасной сети и полная сеть для более безопасной сети. На промежуточном этапе работы оптимальная структура связи соответствует усиленному кольцу («колесо» с радиальными звеньями и связками хорд Рисунок 1.5.

1.2 Понятие цифрового СМИ в сети распространения информации

Люди используют инструменты, а инструменты, которые возможно использовать для общения на расстоянии, во времени и с большим количеством людей одновременно, чем мы могли бы осуществить своим собственным голосом и телом, являются «медиа». Хотя определение может включать в себя межличностные и немассовые средства информирования, такие как телефон, в обычном использовании есть предположение, что «средства массовой информации» предназначены для общения более чем с одним человеком. Традиционные примеры включают книги, журналы, газеты, кино, радио и телевидение.

Однако если переместить радио в интернет, это можно назвать цифровым медиа. А если переместить газету в планшет, то это может считаться цифровым носителем. Проблема с соблюдением этого определения заключается в том, что оно упускает два важных элемента, которые стали возможными благодаря комбинации компьютеров, программного обеспечения и сетей: интерактивность и формирование групп.

Интерактивность стала возможной, потому что большинство компьютерных сетей являются двунаправленными и адресуемыми. Другими словами, возможно указать, куда отправлять сообщение, и сразу же получить ответное сообщение. Эта функция встроена в телефон, но большинство средств массовой информации являются односторонними или широковещательными. Они спроектированы так, чтобы доставлять одно и то же сообщение сразу нескольким людям, но не предоставляют никаких ответных сообщений. Сети цифрового мультимедиа отличаются - возможно отправлять одно и то же сообщение многим людям, например, Netflix, потоковое радио или просто простая веб-страница, но также возможно и взаимодействовать с этими источниками, начиная от второстепенных элементов (выбирая шоу и оценивая их на Netflix) и до основным компонентов (размещение фотографий и комментариев к фотографиям других людей в Instagram).

Вторая уникальная особенность сетевых цифровых носителей состоит в том, что, поскольку они основаны на программном обеспечении, люди,

участвующие в сети, могут объединяться в произвольные группы. Это наиболее очевидно видно в Facebook, где можно легко и быстро создать новую группу по любой теме. Эти «группообразующие» сети имеют значение, поскольку они помогают координировать, общаться и сотрудничать в больших и малых темах.

Таким образом, наиболее важной частью цифрового мультимедиа является не просто преобразование обычных мультимедиа в цифровые форматы. Самое важное - это использовать те возможности, которые представляют из себя интерактивность, а, следовательно, получение обратной связи, а также связность, которая использует все возможности каналов распространения информации [22].

1.3 Отличия взаимодействия СМИ и межличностным общением

Под массовой коммуникацией понимается передача сообщения или информации большой аудитории. При этом есть такое понятие как медиа эффект.

Процессы и продукты влияния средств массовой информации воздействуют непосредственно на цели отдельных лиц и на цели общества и учреждений, а также косвенно на цели через другие механизмы. Эти эффекты могут быть преднамеренными или непреднамеренными как со стороны отправителей мультимедиа, так и целевых получателей. Они могут быть проявлены или скрыты от естественного наблюдения. Они постоянны и непостоянны. И конечно, они формируются не только влиянием СМИ, но и в совокупности других факторов, которые действуют совместно с этим влиянием.

При этом само межличностное общение - это передача сообщения от одного человека другому. Его также можно назвать обменом сообщениями, поскольку обратная связь в основном немедленная. Если обозначить тезисно то:

Сходства:

- Обе формы общения включают отправителя, получателя и носитель;
- Существует вероятность обратной связи в обоих случаях;
- Отсутствие обратной связи подразумевает сбой в коммуникации.

Различия:

- В массовой коммуникации получатели неизвестны друг другу. Но когда общение является межличностным, отправитель и получатель знают друг друга.

- Даже если один или два из большой аудитории не могут понять или справиться с передачей информации, коммуникация остается непрерывной во время массовой коммуникации. Но в случае межличностного общения, если один из отправителей и получателей не может понять сообщение, это приводит к сбою связи.

- Для массовой коммуникации требуется больше средств массовой информации (печатная, электронная или в Интернете), в то время как для

межличностного общения такие средства массовой информации практически не нужны.

1.3.1 Соотношения между коммуникацией с помощью СМИ и межличностной коммуникацией

Связь между средствами массовой информации и межличностным общением может быть осмыслена по-разному [23]. Согласно конкурентной позиции, это две функционально эквивалентные формы общения, которые независимо влияют на пользователей СМИ и участников дискуссии. Следовательно, основной вопрос в том, какие формы общения преобладают.

Исследования, включая сетевой анализ, подтвердили, что межличностное общение во многих отношениях более эффективно, чем массовое общение. Хотя средства массовой информации заменили межличностное общение как важнейшее средство информации во многих областях, когнитивные и убедительные эффекты по-прежнему в основном связаны с межличностным общением. Например, отношение людей к темам средств массовой информации обсуждается только в личной беседе [24], и их решения относительно политических партий и кандидатов основываются преимущественно на предпочтениях в конкретной среде, а не на логике средств массовой информации [25].

Согласно комплементарному положению, межличностное общение и общение в СМИ являются функционально дополняемыми каналами связи. Эта концепция не ставит в приоритет эмпирический вопрос о том, какие каналы связи более активны на соответствующем уровне, а теоретически оценивает потенциальную активность канала на его собственном уровне эффективности.

В соответствии с этим, новейшая теория распространения утверждает, что средства массовой информации охватывают больше людей одновременно, создавая, как правило, основу для распространения информации, в то время как межличностное общение легче преодолевает барьеры выбора, сильнее воздействуя на мнения и взгляды. Кроме того, исследования эффективности кампаний в области здравоохранения показывают, что сообщения в средствах массовой информации достигают полного эффекта только при взаимодействии с межличностным общением. Они также видят потенциальную эффективность сообщений средств массовой информации в распространении тем среди широких слоев населения и в передаче знаний людям используя их же взаимодействие друг с другом. Однако отношение к той или иной теме, связанной со здоровьем, и изменение поведения происходит только в ходе разговоров в личных сетях.

В то время как конкурентная и комплементарная позиции в основном обеспечивают различные интерпретации результатов средств массовой информации и межличностных коммуникационных эффектов, позиция взаимодействия преследует другой подход. Согласно этой точке зрения, две формы общения действуют не независимо, а в сложном взаимодействии,

которое заслуживает изучения. В то время как основная функция средств массовой информации для межличностного общения кажется неоспоримой - освещение ими различных информационных поводов, обеспечивает основу для разговоров по социально значимым темам [26]. С одной стороны, межличностное общение может усиливать медиа-эффекты, передавая информацию, распространяемую медиа, людям, которые не имеют непосредственных контактов с ним. С другой стороны, оно также может служить фильтром для медиаэффектов. В этом случае межличностное общение смягчает медиа-эффекты или даже подавляет их, потому что участники беседы дополняют, интерпретируют и подвергают сомнению медиа-контент с помощью своих собственных (или чужих) фрагментов информации.

Проводимые исследования действительно подтверждают тезисы усилителей и фильтров. Однако у них недостаточная эмпирическая основа. Например, исследования по установлению повестки дня исследуют медиа-контент СМИ, но не разговорный контент личных сетей, в то время как групповые обсуждения отдельных медиа-сообщений не позволяют утверждать о полном содержании полученных медиа-сообщений. Обобщаемые утверждения о совпадении средств массовой информации и межличностного общения и, следовательно, о роли межличностного общения в процессе воздействия средств массовой информации потребуют широкого и систематического анализа контента в двух формах общения.

1.3.2 Инициализация межличностного общения с помощью СМИ

Результаты исследований указывают на то, что медиа-контент прочно закрепился в межличностном общении. Например, обмен с информацией с другими людьми часто называется, как причина для использования медиа. Кроме того, косвенные эмпирические данные для создания общественного дискурса в средствах массовой информации были предоставлены в ходе исследования по установлению повестки дня путем сопоставления повесток дня пользователей СМИ и тех, кто не пользуется ими. Согласно этому исследованию, актуальная повестка дня тех, кто не взаимодействует со СМИ идет с задержкой. Причина может заключаться в том, что средства массовой информации побуждают получателей обмениваться мнениями по темам, о которых сообщают людям из их окружения, которые не имеют доступа к отчетам. Прямое эмпирическое доказательство того, что средства массовой информации инициировали межличностное общение, представлено в исследованиях, которые показывают, как часто респонденты обмениваются мнениями о содержании средств массовой информации, с другими. Также это было отмечено в исследованиях, которые находят корреляции между интенсивностью использования средств массовой информации и количеством участников в политических беседах. В экспериментах можно было доказать, что освещение новостей повышало количество участников в подобных беседах.

Например, ежедневное взаимодействие со СМИ повышает желание людей участвовать в политических дискуссиях с другими людьми. Роль медиа-контента в разговорах также непосредственно наблюдалась в некоторых исследованиях. Наблюдательные исследования варьируются по спектру от качественных откровенных наблюдений за семейными разговорами, проводимыми перед телевизором в середине коллективного просмотра, в личном пространстве [27], или за беседами с друзьями в лаборатории [28], и до количественных скрытых наблюдений социальных групп в публичном пространстве.

Последние предоставляют особенно веские доказательства того, что мы интегрируем разнообразный медиа-контент в наши ежедневные разговоры; однако из-за количества необходимых усилий и этических соображений такие наблюдения встречаются крайне редко [29]. Будет ли медиа-контент инициировать межличностное общение, зависит от характеристик участников и их собеседников, а также от особенностей освещения в СМИ [30].

Участники беседы преимущественно представляют контент из средств массовой информации, которые они часто используют, и темы, которые им очень интересны. Социально-демографические характеристики, такие как пол и возраст, не влияют на частоту передаваемого медиа-контента, однако они влияют на выбор партнера и место для данного разговора. Мужчины обычно обмениваются медиа-контентом на рабочем месте с коллегами, женщины и пожилые люди, как правило, делают это дома, с партнерами, молодежь, как правило, делают это с друзьями. Элементы сообщения, которые поддерживают последующее общение, являются факторами, известными в новостных исследованиях как «персонализация» и «потенциальная возможность конфликта». Более того, диффузионные исследования предполагают, что темы освещения в СМИ инициируют межличностное общение не только чаще, но и быстрее, если в медиа-изображении присутствуют эти факторы или если они успешно воспринимаются получателями. Другими словами, чем ближе освещаемая тема получателю информации, или чем больше эмоций она у него вызывает, тем выше вероятность, что эта информация будет передана ближайшему круг общения получателя.

1.3.3 Влияние актуальности на процесс передачи информации

Хотя средства массовой информации предоставляют темы для межличностного общения, они не определяют, какую актуальность мы им даем в наших социальных сетях и что мы знаем о них. Исследования по установлению повестки дня на совокупном уровне подразумевают, что люди действительно передают актуальные повестки дня, определяемые средствами массовой информации, в значительной степени, не меняя их. Например, эти исследования показали, в частности, что повестки дня СМИ влияют сначала на повестку дня лидеров мнений, а через несколько дней - на остальное население, и что даже повестка дня пользователей, не участвующих в медиа, следует

повестке дня СМИ, хотя и с некоторой задержкой. Однако незначительное влияние индивидуально используемого медиа-контента на актуальные повестки дня получателей может указывать на то, что актуальность тем, о которых сообщают СМИ, не обсуждается, пока они не перейдут в личные разговоры. Это предположение подтверждается исследованиями, в которых записано поведение при разговоре [31] или даже актуальными повестками дня людей из социальных сетей респондентов [24]. Эти исследования показали, что люди, которые обмениваются мнениями с другими о темах СМИ, считают эти темы более важными, чем люди, которые не говорят о них другим, и что тематические повестки дня респондентов могут быть намного лучше объяснены актуальными повестками дня их социальной сети. партнеров, чем по актуальным повесткам СМИ, которые они использовали.

Кроме того, анализ ситуаций коллективного просмотра, обсуждения в фокус-группах и подробные интервью показали, что в ходе бесед медиа-информация не просто передается другим без каких-либо изменений. Согласно этим результатам, участники беседы преобразуют медиа-контент в межличностное общение, дополняя его индивидуальным личным опытом или реализуя аспекты представления подробных аргументов. Обычно происходит, что межличностное общение следует за освещением в СМИ, особенно в случаях кумуляции и созвучия, то есть когда темы одинаково подчеркиваются и изображаются различными средствами массовой информации, а также в случае ненавязчивых вопросов, то есть когда темы выходят за рамки личного опыта получателя. Можно ожидать незначительного влияния на межличностное общение, если различные средства массовой информации подчеркивают одну конкретную тему в разной степени, если изображение противоречиво, фрагментарно или неоднозначно и если темы тесно связаны с элементами собственной реальности получателя, такими как образование и здоровье.

Причиной, по которой люди лучше воспринимают темы межличностного общения в по сравнению с темами, инициированных освещением в СМИ, может быть их личная озабоченность в сочетании с неоднозначной информационной основой и когнитивными ограничениями. Тем не менее, обмен мнениями внутри группы может способствовать пониманию и (правильному) воспоминанию медиа-контента - и, следовательно, образовательного эффекта и эффекта получения знаний средствами массовой информации. Эти результаты подтверждаются исследованиями, в которых говорится, что получатели, которые, по их собственному признанию, часто обсуждали новости с другими, лучше запоминают контент [32]. Сопоставимые результаты доступны для кампаний связанных со здоровьем. Например, получатели, которые обмениваются мнениями по теме, относящейся к тематике (например, наркотики), с другими членами своих социальных сетей, запоминают данную кампанию (скажем, кампанию против наркотиков на телевидении) необычайно хорошо [33]. Лабораторные эксперименты, индуцирующие последующее общение [26], дают доказательства разговоров

как возможной причины улучшения производительности памяти. Тот факт, что другие участники беседы дополняют или исправляют полузабытую или неправильно запомненную информацию из средств массовой информации, объясняет усиление эффектов медийного обучения посредством межличностного общения. Межличностное общение и даже его перспективы могут привести к более интенсивному и тщательному анализу сообщений СМИ по обсуждаемым темам.

1.3.4 Влияние новых средств массовой информации на межличностное общение

Новые формы средств массовой информации и межличностное общение - промежуточные формы, такие как «массовое общение» [34], - развивались вместе с «новыми средствами массовой информации». Прежнее статическое разделение СМИ и те кто с ними взаимодействует, уже неактуальны, и прежние спецификации, такие как отсутствующие параметры обратной связи, более не являются универсально применимыми. Интернет позволяет обычному человеку без особых усилий обращаться к журналистам (даже профессиональным журналистам) за отзывами о своих произведениях. С одной стороны, тот факт, что грань между средствами массовой информации и межличностным общением становится все более размытой, затрудняет изучение влияния содержания средств массовой информации на межличностное общение. С другой стороны, новые средства массовой информации способствуют исследованию таких эффектов, поскольку межличностное общение в режиме онлайн, в отличие от личного общения, становится явным – его возможно отследить к примеру, по личной переписке. Следовательно, значительно легче понять и зафиксировать распространение контента СМИ в социальных сетях [35] или в межличностном общении, стимулируемом медиа [36].

Кроме того, новые коммуникационные технологии влияют на то, как люди общаются в своей социальной среде. Мобильный Интернет обеспечивает различные каналы связи для людей в современном информационном обществе, позволяя обмениваться информацией не только в небольших, но и в более крупных группах и, таким образом, оказывать влияние на межличностное общение. В результате межличностное общение, опосредованное онлайн, отличается от личного общения в пяти основных аспектах [37].

Во-первых, информация в основном сообщается в письменной форме и предоставляет участникам больше времени на обдумывание либо для создания своих собственных сообщений, либо для реакции на сообщения других. В отличие от сообщений, передаваемых в устной форме, сообщения, которыми обмениваются другие пользователи в сети, не меняются, поскольку они фиксируются в письменной форме.

Во-вторых, поскольку отправители часто формулируют свои сообщения в условиях ограниченного социального присутствия, некоторые из них

используют онлайн-платформы в качестве дневников (блогов). Следовательно, личные сообщения часто достигают своих слабых связей.

В-третьих, обновление статуса, сообщения и твиты обычно не предназначены для конкретных получателей. Это создает расширенную самопрезентацию, а также увеличивает шансы на получение социальной поддержки в Интернете из-за большего числа получателей. В-четвертых, другой фактор заключается в том, что увеличение аудитории ведет к расширению обмена контентом, который учитывает интересы получателей.

В-пятых, поскольку люди имеют возможность становиться анонимными в Интернете, вероятность дискуссий по спорным и деликатным темам возрастает. Однако, раскрывая свою личность, собеседники сталкиваются с еще большим социальным давлением и с большей вероятностью адаптируются к мнению большинства в своей виртуальной группе.

Проявление межличностного общения в Интернете не должно заставлять делать поспешный вывод о том, что межличностное общение через Интернет само по себе более актуально, чем личное общение, и, как правило, его можно заменить в качестве объекта исследования. Личные беседы не стали более редкими в эпоху Интернета; по крайней мере, они остаются наиболее важным элементом социализации в частной сфере [38]. Недавно добавленные каналы не заменяют общение лицом к лицу, однако они приводят к диверсификации межличностного общения.

В настоящее время мы устанавливаем социальные контакты и поддерживаем эти связи также с помощью мобильных телефонов или систем мгновенных сообщений. Мы также делимся информацией по электронной почте или в социальных сетях. Однако использование новых, технически опосредованных форм межличностного общения имеет последствия для нашей жизни. Новые сети могут появляться в виртуальных сообществах, и их слабые связи выходят далеко за рамки реальных. Кроме того, устанавливаются новые каналы и способы связи, а старые нормы разрушаются. Ожидание постоянного присутствия и поддержание постоянной связи с партнерами по общению являются яркими тому примерами. Использование новых коммуникационных технологий может также повлиять на наше взаимодействие с другими людьми из нашего социального окружения, под влиянием средств массовой информации [39]. Таким образом, их эффекты могут быть либо разделительными - например, когда транслируемые темы находятся за пределами каналов регулярного взаимодействия, либо отвлекать внимание от ситуации общения, например, когда они предоставляют темы для разговора лицом к лицу или включены в такие темы.

1.4.5 Влияние СМИ на поведение индивидов

Существует несколько подходов организовать все медиа-эффекты, СМИ на человека. Есть две характеристики эффектов, которые особенно полезны. Первая характеристика — это тип эффекта, например, влияет ли этот эффект на

поведение человека, его отношение, эмоции и т. д. Вторая это то, как СМИ оказывают свое влияние на человека. Когда эти характеристики рассматриваются вместе, возможно создать матрицу, которая имеет достаточно категорий, чтобы помочь организовать все эти эффекты.

Есть шесть типов воздействия на людей. Они отличаются в зависимости от характера подверженного влиянию лица или характера переживаний эффекта. Эти шесть типов - познание, вера, отношение, влияние, физиология и поведение. Все исследования медиа-эффектов на индивидуальном уровне изучают, как медиа оказывают влияние на один или несколько из этих шести типов.

Познавательный эффект медиа возникает, когда медиа влияет на психические процессы человека или продукт этих психических процессов. Познавательный эффект легче всего задокументировать, - это получение фактической информации из сообщений средств массовой информации, в частности из книг, газет, телевизионных новостей и информационных сайтов. Человеческий разум может усвоить эту информацию в процессе запоминания. Однако разум может сделать гораздо больше, чем просто запомнить, он может преобразовать информацию в знания. Это преобразование информации может принимать форму логических шаблонов в сообщениях мультимедиа. Человеческий разум также может группировать медиа-сообщения различными способами, чтобы создавать в свою очередь новые значения. Он может обобщать знания и за пределами сообщений СМИ, чтобы генерировать принципы о реальной жизни. Все эти умственные действия оказывают познавательное воздействие на людей.

Убеждения были определены как познания о вероятности того, что объект или событие связано с данным атрибутом. Проще говоря, убеждения - это вера в то, что что-то реально или верно. Средства массовой информации постоянно создают и формируют наши убеждения, показывая нам большую картины мира, чем мы можем увидеть самостоятельно. Никто из нас никогда не встречал Альберта Энштейна, но мы все считаем, что он существовал и был одним из величайших физиков и ученых, потому что о нем мы читали в учебниках, биографиях и научных трудах. Каждый из нас верит в существование очень многих вещей, которые мы никогда не видели непосредственно в нашей реальной жизни так как многие из этих убеждений пришли из сообщений средств массовой информации.

Отношение — это суждение о чем-то. Например, люди видят персонажа в фильме и оценивают его привлекательность, статус героя, симпатичность и так далее. Когда в средствах массовой информации также публикуются истории о людях, событиях, проблемах и вообще в реальном мире. Эти истории часто вызывают у нас необходимость делать собственные суждения о спорных вопросах, политических кандидатах, рекламируемых продуктах и тому подобном.

Влияние также относится к чувствам, которые испытывают люди. Оно включает в себя эмоции и настроения. Средства массовой информации могут

вызывать эмоции, особенно страх, похоть, гнев и смех. Средства массовой информации также предоставляют людям множество возможностей управлять своим настроение. Когда мы чувствуем стресс от всех проблем в нашей реальной жизни, мы можем расслабиться, слушая музыку, забыть о наших проблемах, смотря телевизор, или потерять себя вовремя видеоигры.

Физиологический эффект - это автоматическая реакция организма. Реакция организма может быть либо чисто непроизвольной (например, расширение зрачка, кровяное давление, гальваническая реакция кожи), либо квазиавтоматической (частота сердечных сокращений, половые реакции). Например, когда люди смотрят боевик / приключенческий фильм, их сердцебиение и кровяное давление обычно повышаются. Их мышцы напряжены, а ладони потеют. Они испытывают реакцию «бей или беги», которая была жестко запрограммирована в мозг человека. Угрозы вызывают внимание, и тело готовится к борьбе с хищником или бегству. Этот эффект борьбы или бегства позволил человечеству выжить в течение тысячелетий

Поведение, как правило, определяется как явные действия человека. Исследователи медиа-эффектов провели множество экспериментов, в которых они наблюдают за поведением людей в СМИ, чтобы увидеть, какие медиа они используют и как они их используют. Во время этих экспериментов они также знакомят людей с конкретными сообщениями средств массовой информации, а затем наблюдают за их последующим поведением в отношении таких вещей, как агрессия, использование рекламируемых продуктов и обсуждение политических вопросов.

1.4.6 Действия вызываемые медиа эффектами

Когда какой-либо из шести типов эффектов возникает у человека, необходимо определить, возник ли он под влиянием какого-то события из средств массовой информации. Если входе исследования мы приходим к выводу, что на эффект повлияли средства массовой информации, то получаем эффект от средств массовой информации. Это не означает, что СМИ были единственной причиной такого рода эффекта, вместо этого мы имеем в виду, что они сыграли какую-то роль в достижении этого эффекта.

Эти процессы медиа влияния представляют из себя получение, инициализация, изменение и усиление. Первые два из этих процессов влияют на непосредственные эффекты, которые проявятся либо во время воздействия, либо сразу после него. Третий, изменение, имеет признаки, которые могут проявляться сразу во время воздействия в качестве немедленного эффект, но также есть и другие особенности, которые могут проявиться намного позже. И четвертый процесс требует много времени чтобы проявиться.

Получение. Каждое медиа-сообщение состоит из элементов, и во время воздействия этих сообщений отдельные лица получают и сохраняют некоторые из этих элементов. Сообщения включают в себя такие вещи, как факты, изображения, звуки, отношение ученого к чему-либо, описание

последовательности событий и так далее. Во время показа в СМИ человек мог обратить внимание на определенные элементы в сообщении и сохранить их в своей памяти. Это немедленный эффект, потому что элемент фиксируется в памяти во время показа сообщения. Эта память может длиться несколько секунд или несколько лет, но не то, как долго длится память, определяет, является ли эффект немедленным или нет – определяющим является тот момент времени, когда впервые проявился эффект.

Функция получения применима ко всем типам эффектов, за исключением физиологии, где медиа-сообщения не способны создать физиологические изменения в человеке. Люди получают информацию и хранят ее в своих структурах памяти. Люди также могут приобретать убеждения, отношения, эмоциональную информацию и поведенческие последовательности таким же образом, просто используя навык запоминания. Со всеми этими типами эффектов средства массовой информации создают в уме человека то, чего не было момента передачи информации. Можно утверждать, что все эти эффекты являются по существу когнитивными, поскольку все они требуют использования познавательного навыка запоминания и сохранения информации в памяти человека.

Инициализация. Во время воздействия средств массовой информации СМИ могут активировать то, что уже существует в человеке. Эффект инициализации применим ко всем шести категориям эффектов. Медиа-сообщение может активировать отклик о ранее изученной информации, отзыв уже существующего отношения или убеждения, эмоций, физиологической реакции или ранее изученной последовательности поведения.

Мультимедиа также может запускать процесс, который заставляет человека выполнять многоэтапную задачу. Например, когда люди читают новости о политических кандидатах, о которых они никогда раньше не слышали, они не имеют никакого отношения к этому кандидату. Во время этого освещения люди могут принять информацию из новостной ленты, сравнить ее со своими стандартами для политических кандидатов и создать отношение. Этот процесс отличается от простого приобретения, потому что человек не запоминает чье-то отношение, представленное в средствах массовой информации, а вместо этого проходит процесс конструирования его или ее собственного отношения; в этом случае элемент медиа-сообщения вызвал у человека необходимость выстраивания нового отношения.

Средства массовой информации также могут инициировать процесс реконструкции. Медиа-сообщение может содержать информацию, которая не соответствует существующей структуре знаний человека, поэтому человек должен что-то сделать, чтобы включить новую информацию в свою существующую структуру знаний. Например, допустим, что у человека очень благоприятное отношение к конкретной марке молока, но затем он получает сообщение из СМИ, в котором представлены факты о производстве молока с нарушением технологии и эта новая информация может вызвать переоценку его ранее положительного отношения.

Изменение. Во время воздействия СМИ могут изменить то, что уже присутствует в человеке. Процесс изменения работает со всеми типами эффектов. Сообщения в СМИ могут изменить структуру знаний человека с добавлением новых фактов. Вера может быть изменена, когда средства массовой информации представляют факт, раскрывающий, что существующая вера человека была ошибочной. Средства массовой информации могут изменять индивидуальные стандарты при построении отношений. У людей, которые постоянно подвергаются воздействию элементов историй ужасов и насилия, постепенно стирается их естественная реакция «бей или беги». Постепенно изменяя контент, СМИ могут изменить настроение человека.

Изменение может проявиться немедленно (то есть во время воздействия или сразу после воздействия сообщения мультимедиа), или оно может занять много времени. Изменение может быть временным (и исчезать через несколько секунд) или может сохраняться достаточно долго. Большая часть исследований долгосрочных медиа-эффектов основана на предположениях о долгосрочном медиа-влиянии как постепенном процессе формирования. Это своего рода процесс «сообщение за сообщением», который медленно меняет структуру наших знаний. При этом существуют также и «слабые» влияния. Ведь не все сообщения СМИ имеют одинаковое влияние и что не все персонажи в материалах СМИ одинаково влияют на наши убеждения и отношение. Некоторые изображения выделяются, потому что они ненормальны, интенсивны и, следовательно, оставляют более запоминающиеся впечатлениями от просмотра.

Усиление. Посредством повторяющихся воздействий средства массовой информации постепенно увеличивают вес того, что уже есть в сознании человека, тем самым делая это что-то более постоянным и трудным для изменения. Функция усиления применима ко всем шести типам эффектов. Когда в средствах массовой информации постоянно появляются одни и те же люди и события в новостях, структуры знаний отдельных людей об этих людях и событиях становятся более жесткими и с меньшей вероятностью измененяются позже. Когда средства массовой информации представляют те же убеждения и взгляды, уровень комфорта людей с этими убеждениями и отношениями становится настолько сильным, что они не могут их изменить. Когда средства массовой информации представляют сообщения одного и того же типа каждую неделю или каждый день, поведенческие модели воздействия людей становятся более фиксированными и их труднее изменить.

1.4 Методы определения и борьбы с недостоверной информацией

Использование и значение термина поддельные новости развивались с течением времени. Анализ Google Trends этого термина обнаруживает внезапный всплеск популярности во время президентских выборов в США 2016 года а также в последнее время в связи с эпидемией коронавируса (Рисунок 1.6). Хотя первоначально он использовался для обозначения на

ложной и часто сенсационной информации, распространяемой под видом новостных сообщений, этот термин эволюционировал и стал синонимом распространения ложной информации [40]. Фейковые новости, как правило, определяются как «новостная статья, которая является преднамеренно и достоверно ложной» [41, 42] или «информация, представленная как новостное сообщение, которое на самом деле неверно и предназначено для того, чтобы обмануть потребителя, который полагает, что это правда» [43]. , Однако существующие определения являются узкими, ограниченными либо типом информации, либо намерением обмана, и не охватывают более широкую область действия термина на основе его текущего использования. Поэтому возможно определить фейковые новости следующим образом: новостная статья или сообщение, публикуемое и распространяемое через средства массовой информации, несущее ложную информацию независимо от средств и мотивов, стоящих за ней.

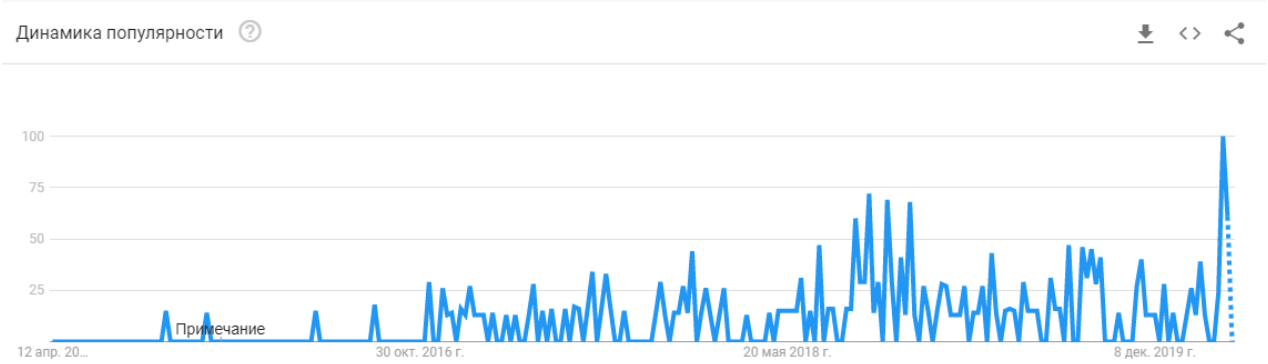

Рисунок 1.6 – Динамика популярности поискового запроса «fake news»

Это определение позволяет нам фиксировать различные типы фальшивых новостей, которые могут быть дифференцированы с помощью средств, используемых для фальсификации информации, таких как сфабрикованный контент (полностью ложный), вводящий в заблуждение контент (вводящее в заблуждение использование информации для постановки проблемы), подталкивающий контент (подлинные источники подражают ложным источникам), манипулируемый контент (подлинная информация или изображения, которыми манипулируют, чтобы обмануть), ложная связь (заголовки, визуальные элементы или подписи, которые не связаны с контентом) и ложный контекст (подлинный контент, предоставленный ложной контекстной информацией). Определение также позволяет нам включать в себя различные типы поддельных новостей, которые определяются по их мотивам или намерениям, например, злонамеренное действие (причинить вред или дурную славу), прибыль (для получения финансовой выгоды за счет увеличения количества просмотров), влияние (для манипулирования общественным мнением), сеять раздор (для создания беспорядка), страсть (для продвижения идеологических предубеждений), развлечения [43]. Мы также можем разделить ложную информацию по умыслу как недопонимание и

дезинформация. Недопонимание относится к непреднамеренному распространению ложной информации, которая может быть результатом искажения информации, вызванных когнитивными искажениями или отсутствием понимания или внимания; а дезинформация относится к ложной информации, созданной и распространяемой конкретно с намерением обмануть [44]. Другим типом информации, которая может быть тесно связана с фальшивыми новостями, является сатира - сатира представляет истории как новости, которые могут быть фактически неверными, но цель состоит не в том, чтобы обмануть, а призвать, высмеять или разоблачить поведение, которое является постыдным, коррумпированным, или иначе «плохим» [45].

Намерение, стоящее за сатирой, кажется достаточно законным, чтобы исключить его из определения, однако [46] действительно включает сатиру как тип фальшивых новостей, когда нет намерения причинять вред, но он может ввести в заблуждение или обмануть людей. Кроме того, упоминается, что существует целый спектр от поддельных до сатирических новостей, которые, как они обнаружили, используются многими поддельными новостными сайтами, которые размещали заявления о недостоверности информации в нижней части своих веб-страниц, чтобы предположить, что они были «сатирическими», даже если в их статьях не было ничего сатирического. Это сделано для того чтобы защитить себя от обвинений в том, что они размещают недостоверную информацию. Таким образом, определение должно включать статьи, которые ложно обозначены как сатира, а также сатирические статьи, которые могут ввести в заблуждение, и исключать другие, которые не попадают в эту область. Кроме того, рассмотреть термины обман и слух, которые тесно связаны с поддельными новостями. Обман считается ложной историей, используемой для маскировки правды, и, согласно традиционному определению, поддельные новости можно рассматривать как форму обмана, обычно распространяемого через новостные агентства [47]. Термин «слух» относится к необоснованным заявлениям, которые распространяются при отсутствии доказательств в их поддержку, что делает их очень похожими на фальшивые новости, с основным отличием в том, что они необязательно представляют из себя ложь, и могут оказаться правдой [45]. Слухи происходят из непроверенных источников, но могут позже быть подтверждены как истинные или ложные, или остаются неразрешенными. Таким образом, определение фейковых новостей может быть воспринято как натуральная ложь и ложные слухи.

Существуют три методологии по тому как определить распространение фейковых новостей. Первый тип - идентификация фальшивых новостей с использованием методов, основанных на контенте, которые классифицируют новости на основе содержания. Второй тип - идентификация с использованием методов обратной связи, которые классифицируют новости на основе ответов пользователей, которые они получают в социальных сетях. Наконец, третий тип - это решения, основанные на вмешательстве, которые предоставляют вычислительные решения для активной идентификации и сдерживания

распространения ложной информации, а также методы для смягчения воздействия ложной информации. Каждая категория далее делится на основе типа существующих методов, как показано в таблице 1.

Таблица 1.1 – Категоризация существующих методов анализа распространения информации

Анализ на основе контента	**Анализ на основе обратной связи**	**Анализ на основе вмешательства**
Методы на основе признаков и меток	Ручной анализ текста	Стратегии смягчения последствий
Лингвистический анализ	Анализ характера распространения	Идентификационные стратегии
Глубокое обучение на основе контента	Анализ временных паттернов	
	Анализ текста ответа	
	Анализ ответа пользователя	

В данной диссертационной работе анализ будет проводиться на основе распространения новостей об отставке Первого президента Республики Казахстан – Назарбаева Н.А. И так как объектами исследований являются статьи в СМИ, следовательно, единственный подходящий способ для анализа фейковых новостей это первый способ – на основе содержания самой новости. Далее будет подробнее рассмотрен именно этот подход.

1.4.1 Анализ информационных сообщений на основе их контента

Далее приводится обзор подхода к обнаружению фейковых новостей, основанных на контенте. В основе обнаружения контента лежит то, что текст в фальшивых новостях отличается от настоящего в некоторых количественных отношениях. Использование языковых подсказок для определения правдивости было впервые мотивировано работой в области прикладной психологии для оценки свидетельств очевидцев [48]. Языковые подсказки можно использовать с помощью традиционных методов, основанных на разработке функций для анализа, методов лингвистики и более продвинутых методов глубокого обучения, которые не требуют разработку функций для анализа контента.

1.4.2 Анализ на основе меток и признаков в сообщении

Методы, основанные на метках и признаках, могут использоваться для того, чтобы отличить поддельное новостное содержимое от истинного новостного содержимого путем разработки набора лингвистических подсказок,

которые являются явными признаками для подтверждения достоверности контента.

Анализ научного контента (SCAN).

Одна из самых ранних работ по изучению использования лингвистических подсказок для обнаружения поддельных новостей была написана Дрисколлом 1994, в которой изучались стенограммы или письменные заявления, сделанные отдельными лицами в ходе уголовного расследования. Чтобы определить достоверность информации, предоставленной подозреваемыми в таких заявлениях, они использовали подход, называемый «Анализ научного контента» (SCAN), который был предложен экспертом по полиграфии Сапиром в 1987 году, основываясь на его опыте с предметами проверок на полиграфе. SCAN состоит из сигналов, связанных с обнаружением обмана. Подсказки включают в себя содержание и структурные критерии, такие как отсутствие связи между абзацами, отсутствие убежденности или памяти, отрицание обвинений, отсутствие и неупорядоченная информация, использование эмоциональных слов, объективных или субъективных слов, местоимений от первого лица, единственного числа и глаголы прошедшего времени. Хотя исследование, проведенное Дрисколлом 1994, выявило положительные результаты в различении истинных и ложных утверждений с использованием SCAN, более поздние работы показали, что они неэффективны, основываясь на более строгих оценках, не обнаружив существенных различий между истинными и сфабрикованными утверждениями, проверенных по этим критериям [49, 50].

Несмотря на то, что SCAN интуитивно понятен, он представляет собой набор субъективных критериев, в которых отсутствуют достаточные обоснования эффективности и, кроме того, этот подход требует использования подготовленных специалистов для анализа утверждений на достоверность, что затрудняет автоматизацию.

Лингвистический набор признаков (LBC).

В попытке уменьшить участие человека в обнаружении фейков одним из новаторских методов автоматического анализа текста стала работа Фуллера в 2009 году. В ней он объединил несколько лингвистических сигналов и ранее предложенных наборов сигналов. Первым использованным набором сигналов был набор Чжоу / Бургун [51], включающий 14 лингвистических сигналов, которые были сочтены эффективными для обнаружения обмана, включая процент местоимений от первого лица, среднюю длину слова, количество глаголов, сенсорное отношение, пространственное и временное соотношение и образы. Второй набор сигналов был получен из конструкций обмана, взятых из теорий обмана [52, 53], которые включали количество предложений и слов, активацию, термины определенности, обобщающие термины, образы и количество глаголов. Третий набор сигналов был набором из 31 сигнала, созданным путем объединения первых двух наборов сигналов вместе с дополнительными сигналами на основе лингвистического анализа и подсчета слов (LIWC) [54], которые использовались в предыдущих исследованиях [55,

56], и включал лексическое разнообразие, модальные глаголы, пассивные глаголы, эмоциональность, термины и избыточность. Чтобы определить относительную важность сигналов, Фуллер использовал три различных классификатора, то есть нейронную сеть, дерево решений и логистическую регрессию, накладывая их на заявления свидетелей в официальных расследованиях. При этом вектор входных признаков в классификатор состоял из нормализованной частоты появления сигналов в тексте. Восемью признаками, полученными при выборе признаков, были местоимения от третьего лица, разнообразие слов, исключительные термины, лексическое разнообразие, модификаторы, количество предложений, количество глаголов и количество слов.

Недостатком использования лингвистических наборов сигналов является отсутствие обобщения для тем, языков и областей. Али и Левин в исследовании 2008 года показали, что лингвистический набор сигналов, разработанный для одной ситуации, может не подходить для другой ситуации из-за языковых вариаций, например, набор сигналов, предназначенный для разговора или полицейского допроса, может значительно отличаться.

Другие варианты.

В более поздних работах были исследованы усовершенствованные наборы сигналов, выявленных вручную, более точно ориентированные на проблему обнаружения поддельных новостей. Рубин в 2016 году проанализировал некоторые особенности текста, такие как количество знаков препинания и настроение текста. В 2015 году предложил различные регулярные выражения для сбора шаблонов фейковых новостей и их опровержений в сообщениях в социальных сетях, которые содержат сообщения с общей тематикой «это неправда», которые формируют шаблон, и «это не соответствует действительности» которые фиксируют опровержение. Они также включали некоторые специфичные для платформы функции, такие как подсчет «хэштегов» и «упоминаний» в постах в Twitter, а также соотношение фейковых новостей и опровержений в кластере постов, имеющих высокое текстовое сходство, то есть постов, в которых обсуждается похожий контент. Другие работы, в которых также были предложены наборы сигналов, разработанные специально для определенных типов платформ и социальных сетей, включают [57, 58, 59] для Twitter, [60] для Википедии, [61] для веб-сайтов с намеренно ложной информацией. Однако необходимо учитывать что, исчерпывающее перечисление шаблонов регулярных выражений является нетривиальной задачей и требует значительных усилий. Кроме того, выявление соответствующих специфических функций для большого разнообразия социальных сетей также является сложной задачей и снижает применимость метода как общего подхода к выявлению фейковой информации в различных источниках.

Если суммировать недостатки и различия в анализе, основанном на лингвистических сигналах и метках, необходимо отметить, что для новой ситуации должен быть разработан новый набор сигналов, что затрудняет

обобщение методов для разных тем и областей. Таким образом, такие подходы предполагают более активное участие человека в процессе разработки, оценки и использования этих сигналов для обнаружения.

1.4.2 Анализ на основе лингвистических особенностей

Хотя ручное составление набора меток интуитивно понятно и интерпретируемо, оно часто зависит от рассматриваемой тематики и не может применяться повсеместно. В попытке сделать модели на основе сигналов более общими, были предложены методы, основанные на лингвистическом анализе. Методы, основанные на лингвистическом анализе, такие как методы, основанные на репликах, могут применяться для различения фальшивых и настоящих новостей, используя различия в стиле письма, языке и настроения. Такие методы не требуют специфичных для задачи, ручных наборов сигналов и полагаются на автоматическое извлечение лингвистических особенностей из текста. Далее рассматриваются три метода лингвистического анализа, которые применяются для обнаружения поддельных новостей.

N-граммы.

Первоначально использование n-граммов состояло в следующем: исследователи строили наборы данных, используя краудсорсинг, который представлял собой заявления людей, которые лгут о своих убеждениях по таким темам, как аборты и смертная казнь. Исследователи хотели определить, чем отличаются тексты и достаточно ли анализа n-граммов, чтобы отличить ложь от правды. Они обучили байесовские и опорные векторные классификаторы (SVM), где в качестве входных данных использовались частотные векторы n-граммов в текстах после обработки, но без удаления стоп-слов. Интересно, что точность классификации составляла около 70% при выявлении лжи людей об их убеждениях и 75% при выявлении лжи об их чувствах. Детальный анализ использования слов показал, что во всех вводящих в заблуждение текстах отсутствуют связи с самим собой («я, друзья, мои»), и доминируют классы слов, связанные с другим человеком («вы, другие, люди»), что свидетельствует о дискомфорте говорящего при отождествлении себя с ложными утверждениями. Кроме того, было обнаружено, что слова, относящиеся к «определенности», являются доминирующими в вводящих в заблуждение текстах, что, вероятно, объясняется необходимостью для говорящего явно использовать слова, связанные с истиной, чтобы сделать их ложные утверждения более правдоподобными. Дальнейшие исследования [62, 63] предоставили аналогичный n-граммовый анализ для классификации вводящих в заблуждение обзоров, созданных сотрудниками Amazon Mechanical Turk, которых попросили создать поддельные положительные отзывы об отелях [63] и поддельные отрицательные отзывы об отелях [62]. Их анализ показал, что поддельные отзывы содержали меньше пространственных слов (местоположение, этаж, маленькие), потому что человек на самом деле не посещал гостиницу и имел меньше пространственных деталей, доступных для

обзора, а также было обнаружено, что слова положительного настроения были преувеличены в положительных поддельных отзывах по сравнению с их истинными аналогами. Подобное преувеличение было замечено в словах с отрицательным настроением в поддельных отрицательных отзывах.

Однако будучи упрощенным подходом, использование одних n-граммов не может полностью охватить тонкую лингвистическую информацию, присутствующую в различных стилях написания поддельных новостей.

Анализ структуры текста по количеству частей речи.

Помимо методов анализа, основанных на словах, таких как n-граммы, для формирования языковых характеристик текстов используются еще и синтаксические функции, такие как теги частей речи (POS). POS-теги получают, помечая каждое слово в предложении в соответствии с его синтаксической функцией, например, существительные, местоимение, прилагательные. В нескольких работах было установлено, что распределение частот POS-тегов тесно связано с жанром рассматриваемого текста, например, медицинские консультации, заседания комитетов и проповеди, каждая из которых имеет свой характерный паттерн [64, 65]. В 2011 году было рассмотрено, существует ли эта разница в распределении POS-тегов при анализе достоверности текста. Они обучили классификатор SVM, используя относительные частоты POS-тегов текстов в качестве функций в наборе данных, содержащих поддельные обзоры. При этом была достигнута лучшая классификация с использованием подхода n-грамм, но, тем не менее, было установлено, что подход с использованием POS-меток является сильной базой, превосходящей лучшую оценку человека. Качественный анализ показал, что веса, усвоенные классификатором, в значительной степени согласуются с выводами существующих теорий обманного письма, таких как [65], которые предполагают связь обманных мнений с образным письмом, включающим больше глаголов, наречий, местоимений, и правдивые мнения с информативным письмом, включающему больше существительных, прилагательных, предлогов и соединительных союзов.

Необходимо учитывать, что использование лишь одних POS-тегов дает только синтаксическую структуру текста и что является не столь информативным, чем при использовании подходов, основанных на анализе слов. Они в свою очередь предоставляют больше информации, включая стиль письма, что может быть отражено в эмоциональности анализируемого текста.

Вероятностная контекстно-свободная грамматика.

В более поздних научных работах были рассмотрены глубокие синтаксические особенности, полученные из деревьев вероятностных контекстно-свободных грамматик (PCFG) [66]. Дерево контекстно-свободной грамматики (CFG) представляет грамматическую структуру предложения с конечными узлами, представляющими слова, и промежуточными узлами, представляющими синтаксические составляющие, такие как глагол, существительная фраза и т. Д. В зависимости от неоднозначности конструкции языка, предложение может иметь несколько синтаксических представлений.

PCFG позволяет устранять неоднозначность, связывая вероятность с каждым деревом, где вероятность дерева является произведением вероятностей всех правил производства в дереве. Производственное правило представляется следующим образом: $A \rightarrow \alpha$, где $A \in V$ и $\alpha \in (V \cup T)$, где V является набором промежуточных узлов, а T это набор конечных узлов.

Фенгом в 2012 году было рассмотрено использование PCFG для кодирования более глубоких синтаксических функций для обнаружения обмана. В частности, он предложил четыре варианта кодирования правил производства в качестве функций. Первый вариант включает в себя только те производственные правила из данных, которые не содержат терминальных узлов. Второй вариант включает в себя все производственные правила, полученные из набора данных. Третий и четвертый варианты модифицируют производственные правила для включения узлов-прародителей (то есть родительского узла A в правиле $A \rightarrow \alpha$), с правилами и без правил в отношении терминальных узлов соответственно. Вектор признаков строится с использованием счетчиков (нормализованной частоты) производственных правил в тексте. В ходе исследования был обучен классификатор SVM и удалось обнаружить, что функции PCFG, используемые с функциями n-граммов, более полезны, чем теги POS с n-граммами, при анализе поддельных текстов из обзоров отелей и наборов данных по обнаружению лжи [67, 63]. Было отмечено, что из четырех вариантов производственных правил те, в которые включены терминальные узлы, были более эффективными, поскольку они содержали информацию о структуре текста по частям речи в дополнение к информации о синтаксисе. Качественный анализ синтаксических составляющих, основанный на весах классификатора, показал, что использование некоторых синтаксических составляющих, таких как пассивный залог, глагольные фразы и сложноподчиненные предложения, чаще встречается в поддельных текстах нежели в настоящих.

Этот подход в анализе может использоваться для извлечения синтаксических особенностей из предложения, но недостаточно эффективен для сбора контекстно-зависимой информации между предложениями, что ограничивает его эффективность при классификации более длинных поддельных новостных статей или текстов.

Если подводить итог по методам, основанным на лингвистическом анализе то даже при наличии основанных на словах n-грамматических функций в сочетании с более глубокими синтаксическими особенностями деревьев PCFG методы лингвистического анализа, хотя и лучше, чем методы, основанные на метках, все же не позволяют полностью извлечь и использовать богатую семантическую и синтаксическую информацию в контенте. Подход n-грамм прост и не может моделировать более сложные контекстные зависимости в тексте. Анализ синтаксических особенностей, применяемый в одиночку, менее эффективен, чем n-граммы, основанные на отдельных словах, а их совместное использование не может уловить их сложную взаимозависимость.

1.4.3 Методы глубокого обучения на основе контента

Методы глубокого обучения устраняют недостатки методов, основанных на лингвистическом анализе, благодаря автоматическому извлечению признаков, позволяя получать как простые, так и более сложные функции. Методы, основанные на глубоком обучении, продемонстрировали значительные успехи в классификации и анализе текста [68, 69, 70] и являются эффективными способами для анализа признаков недостоверной информации благодаря способности отслеживать более сложные паттерны.

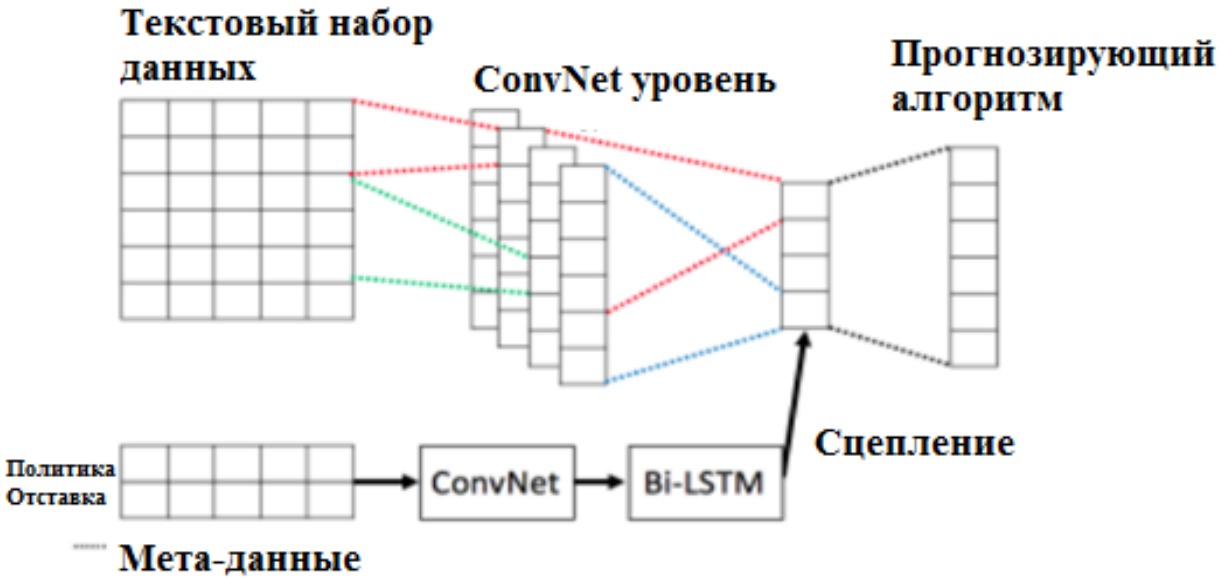

Рисунок 1.7 – Архитектура сверточных нейронных сетей для выявления ложных новостей

Сверточные нейронные сети.

Сверточные нейронные сети (CNN) [71] обычно используются в задачах обработки естественного языка, таких как семантический анализ [70] и классификация текста. Ванг в 2017 году предложил использовать сверточные нейронные сети для обнаружения поддельных новостей на основе контента. Он сформировал набор коротких заявлений, основанных на степени лжи, созданной PolitiFact, признанной организацией, которая занимается разоблачением недостоверной информации. На рисунке 1.7 показана архитектура модели, предложенная в его работе [72].

Модель использует два типа данных на вход: текст заявления и метаданные источника, такие как политическая ориентация, исходное состояние и т. д., Доступные в наборе данных. Ввод текста обрабатывается слоем, который осуществляет встраивание слов для получения непрерывных низкоразмерных представлений для каждого слова в текстовой последовательности. Выходные данные этого слоя обрабатываются

сверточным и максимальным пулами, которые генерируют представление объекта. Точно так же метаданные источника аналогичным образом обрабатываются другим уровнями внедрения и двунаправленным уровнем LSTM для генерации своего окончательного полученного представления признаков. Два представления объединяются и передаются в обученный классификатор сквозным образом с другими уровнями. Ванг в своем исследовании использовал предварительно обученные вложения word2vec [73], чтобы начать встраивание текста.

Известно, что эти вложения слов фиксируют полезные свойства совпадений слов и контекстные свойства, а также семантические отношения между словами, обученные на больших объемах размеченных текстов. В этом исследовании возможно наблюдать лучшую точность обнаружения недостоверной информации по сравнению с использованием SVM и логистической регрессии, а также было обнаружено, что включение метаданных докладчика также является важным. Цянь и соавторы в 2018 году также продемонстрировали улучшение производительности CNN по сравнению с методами, основанными на лингвистическом анализе, такими как LIWC, POS и n-граммный подход, при классификации новостных статей как поддельных или истинных. Кроме того, для обработки более длинных текстов статей был предложен вариант архитектуры CNN, называемый двухуровневой сверточной нейронной сетью (TCNN), который сначала берет среднее значение векторов распределения для слов в предложении, а затем представляет статьи в виде последовательности представлений предложений, предоставляемых в качестве входных данных к сверточным и пулирующим слоям. Цянь и соавторы в 2018 году показали, что вариант TCNN более эффективен при классификации статей, чем CNN.

Рекомендованные архитектуры на основе нейронной сети (RNN) [74] также предлагаются для обнаружения поддельных новостей. RNN обрабатывают вложения слов в текст последовательно, по одному слову / токену за раз, используя на каждом шаге информацию из текущего слова, чтобы обновить его скрытое состояние, в котором собрана информация о предыдущих словах. Окончательное скрытое состояние обычно принимается как представление функции, выделенное RNN для данной входной последовательности. Конкретный вариант, называемый Long Short-Term Memory (LSTM), который облегчает некоторые трудности обучения в RNN, часто используется из-за его способности эффективно фиксировать долгосрочные зависимости в тексте и применяется для обнаружения поддельных новостей, похож на использование сверточных нейронных сетей, в нескольких работах; в то время как в другом варианте LSTM был применен как к заголовку статьи, так и к тексту статьи (основной части), чтобы попытаться классифицировать уровень разногласий между ними для обнаружения обмана [75].

Рекуррентные нейронные сети [74] также используются для выявления поддельных новостей. RNN обрабатывают вложения слов в текст

последовательно, по одному слову / токену за раз, используя на каждом шаге информацию из текущего слова, чтобы обновить его текущее состояние, в котором собрана информация о предыдущих словах. Окончательное состояние обычно принимается как представление функции, выведенное RNN для данной входной последовательности. Один из вариантов, называемый Long Short-Term Memory (LSTM), который облегчает некоторые аспекты обучения в RNN, часто используется из-за его способности эффективно фиксировать сложные зависимости в тексте и применяется для обнаружения недостоверных новостей. По своей сути он похож на использование сверточных нейронных сетей, однако в некоторых случаях варианте LSTM был применен как к заголовку статьи, так и к тексту статьи (основной части), чтобы попытаться классифицировать уровень разногласий между ними для обнаружения обмана [75].

При этом даже с извлечением сложных функций из методов глубокого обучения, обнаружение поддельных новостей остается проблемой, в первую очередь потому, что контент создается так, чтобы он напоминал правду, чтобы обмануть читателей и поэтому без проверки фактов или дополнительной информации зачастую трудно определить достоверность только с помощью анализа текста. Недавняя оценка, сравнивающая различные методы для наборов данных о политических заявлениях и новостях, также подтверждает относительно низкую точность классификации - 63% и 70% для двух наборов данных с использованием только классификации с CNN [76].

1.5 Примеры инструментов отслеживания информации в Казахстане

Рассмотренные технологии анализа средств массовой информации не могут оставаться лишь теорией и поэтому данные подходы активно применяются в коммерческих продуктах. Цели эти продуктов могут быть разные: отслеживание положительных и негативных упоминаний брендов, сбор статистики по количеству упоминаний, а также отслеживание трендов.

В Казахстане активно используются лишь некоторые из этих решений и это связанно с необходимостью анализировать текст на русском языке, что поддерживают не все системы, а также с общей небольшой величиной рынка. Далее будут рассмотрены три главных инструмента по отслеживанию информации в СМИ и социальных сетях, другими словами распределенных системах: Google Alerts, Brand analytics, YouScan, а также будет произведен анализ применения данных методов в рамках казахстанского проекта IMAS.

Google Alerts

«Google Alerts» – это решение предоставленное компанией Google для того чтобы используя все данные, полученные из поисковых запросов исползовать для обнаружения и отслеживания изменений в контенте. При этом пользователю приходит уведомление на почту, когда по указанному ранее поисковому запросу появляются новые результаты. Этом может быть как отдельные веб-сайты, упоминания в СМИ, научные статьи, блоги и т.д. Данное

решение было запущено в 2003 году и до сих пор активно эксплуатируется, что говорит о его востребованности.

Компания Google использует в своей аналитике не только данные из поисковых запросов, но также данные по отдельно найденным пользователям, а также статистику использования своего браузера Google chrome для того чтобы понимать релевантность того или иного контента. Также при сборе данных применяются результаты работы операционной системы android, которая передает данные о месте жительства объекта, маршруте передвижения и т.д., следовательно, на входе у системы Google alerts самый большой объем данных из всех возможных.

Чтобы понять принцип работы этой системы, необходимо понять, как в общем и целом функционирует поисковый движок: поисковый бот Google посещает каждую из возможных веб-страниц, которая подверглась индексации. Оттуда парсится вся доступная информация, и уже этот бот принимает решение, насколько и в каком объеме необходимо передавать эту информацию далее. После чего информация проходит через несколько сотен алгоритмов обработки данных. Сам же сервис оповещений представляет из себя лишь надстройку над основным поисковым движком, которая фиксирует изменения во вновь полученных результатах по сравнению с историческими данными.

При этом существуют несколько параметров, на основе которых Google Alerts будет осуществлять анализ. Например, возможно при передаче атрибута site сузить критерии поиска только до одного источника, а также указать на каком языке необходимо обрабатывать информацию. Последняя особенность действительно сложна в реализации, так как для анализа текста на разных языках, необходимо учитывать их синтаксическую и лексическую составляющую.

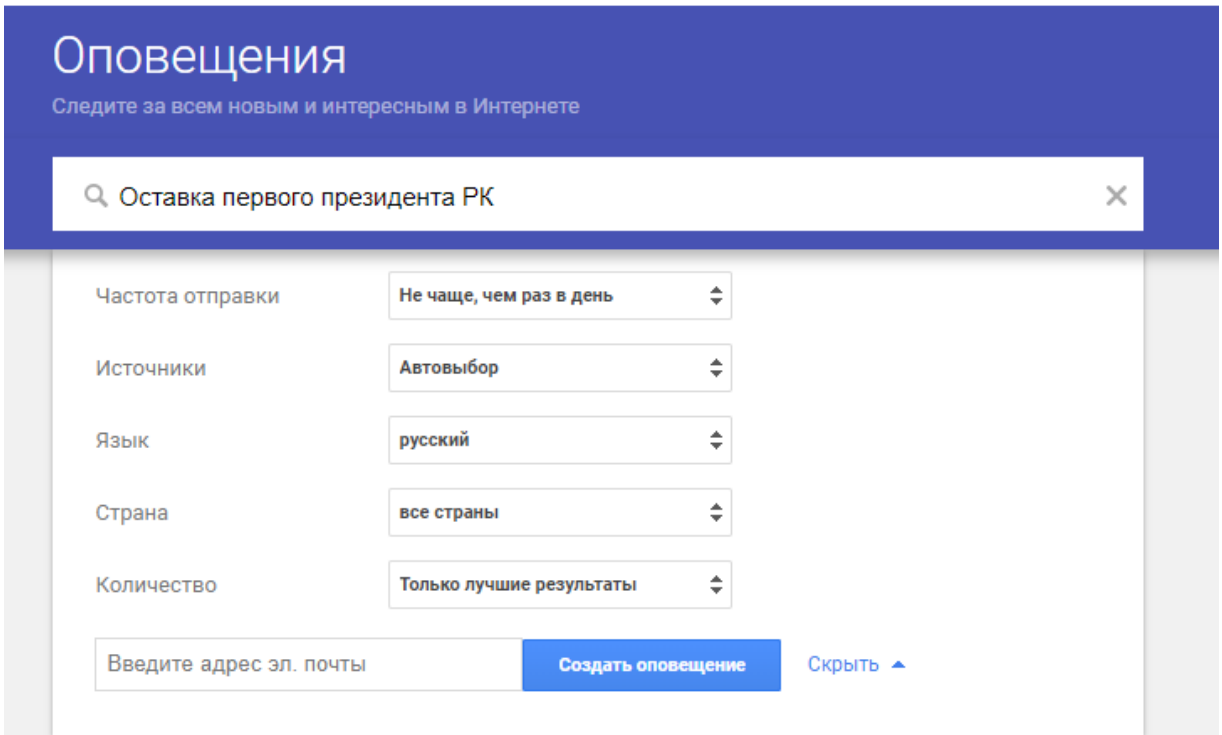

Рисунок 1.8 – Настройка поисковой выдачи в Google Alerts

Brand Analytics

Brand Analytics – это система которая отслеживает упоминания в социальных сетях в реальном времени и определяет их тональность, выявляет тренды, и формирует на основе этой информации отчеты. Инструмент используется больше для отслеживания репутационных рисков для компаний.

Источниками в данном случае выступают как социальные сети: Facebook, Twitter, YouTube, VK, а также множество СМИ. В отличие от Google здесь задействуется не поисковый движок, а отдельные боты для каждого источника, которые парсят весь контент ресурсов, имеющих ценность для целевой аудитории компаний (Рисунок 1.9).

Также отслеживается информация по авторам публикации, такая как пол, возраст, специализация и т.д. Все это заносится в единый справочник и применяется при анализе сообщений на других ресурсов, но от того же автора.

В данной системе ежедневно добавляется около 10 миллионов новых сообщений, а общее количество записей приближается к 5 миллиардам. На основе всей этой собранной информации возможно сформировать осмысленный поток сообщений в зависимости от того что необходимо отследить.

#	Источник	Сообщений	Сообщений %
1	twitter.com	48253	60.52
2	vk.com	19154	24.02
3	livejournal.com	3572	4.48
4	youtube.com	917	1.15
5	news.rambler.ru	523	0.66
6	www.rosbalt.ru	272	0.34
7	elitetrader.ru	232	0.29

Рисунок 1.9 – Количество упоминаний в зависимости от источника информации

Благодаря тому что эта система постоянно занимается отслеживанием изменений на некоторых ресурсах, таких как Twitter данные могут поступать в режиме реального времени, а из некоторых источников, например цифровых СМИ в течение нескольких минут.

Для того чтобы дать количественную оценку интереса аудитории к исследуемым объектам применяется непрерывный анализ уже имеющихся и вновь поступающих данных (Рисунок 1.10).

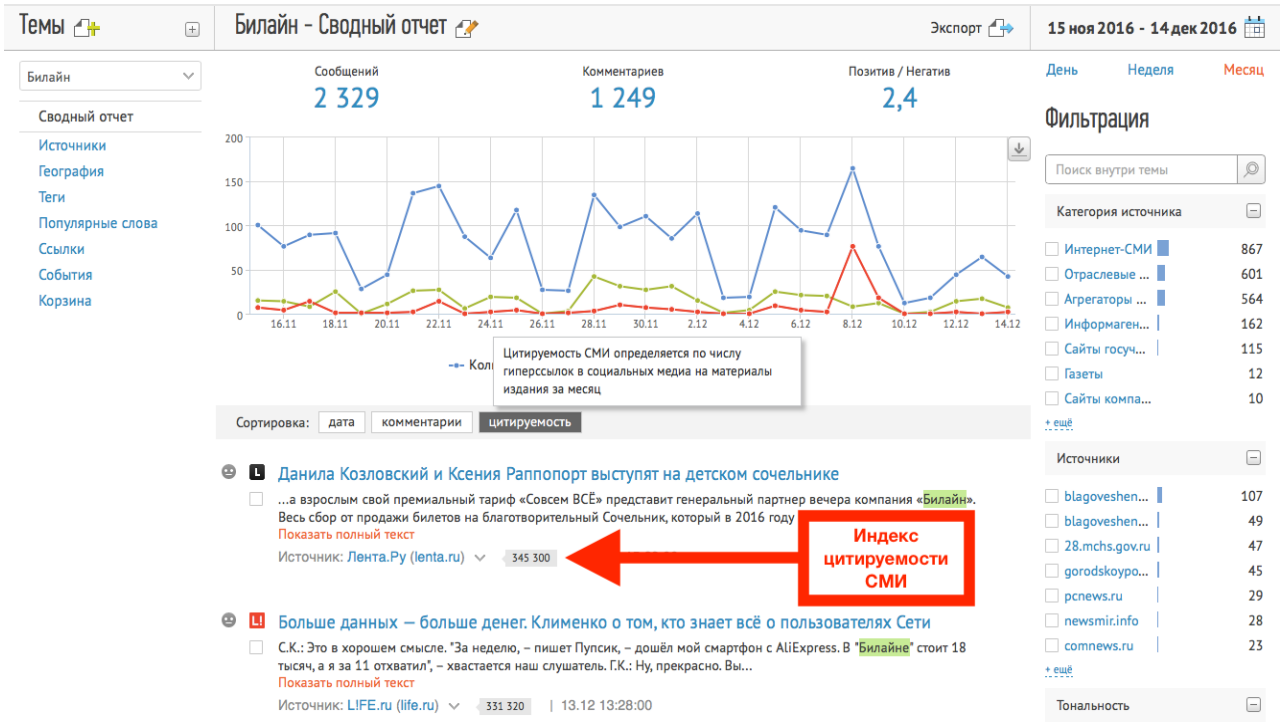

Рисунок 1.10 – Сводный отчет в Brand Analytics по компании Билайн

В качестве методов анализа применяется рассмотренный ранее лингвистический анализ:

- Определяется тональность сообщения на основе ключевых слов;
- Собирается портрет автора на основе его публикаций, географии, пола, возраста и меток в самом тексте;
- Сообщения группируются на основе их схожести между собой и проценте плагиата, а также языка публикации;
- На основе анализа по тегам и меткам, возможно классифицировать текст по отношению к той или иной тематике;
- Как самый практически применимый результат, на основе собранной информации возможно оценить текущий и потенциальный обхват аудитории по распространению информации.

YouScan

Платформа YouScan это еще одна крупная система для отслеживания изменения информации в социальных медиа, которая при этом построена с использованием принципов машинного обучения и применяется для проведения маркетинговых исследований. Так с помощью нее возможно оценить потенциальный обхват аудитории и распространения информации и тем самым правильно спланировать маркетинговый бюджет.

Как результат анализа можно принять:

- возможность отслеживать репутационные угрозы, кризисные ситуации;
- оценить реакцию аудитории на тот или иной продукт компании / новость, путем отслеживания тональности сообщений (Рисунок 1.11);

- в зависимости от того как каждый отдельный человек реагирует на один и тот же контент, возможно сформировать его портрет, вывести привычки, предпочтения, но что самое главное, определить лидера мнений.

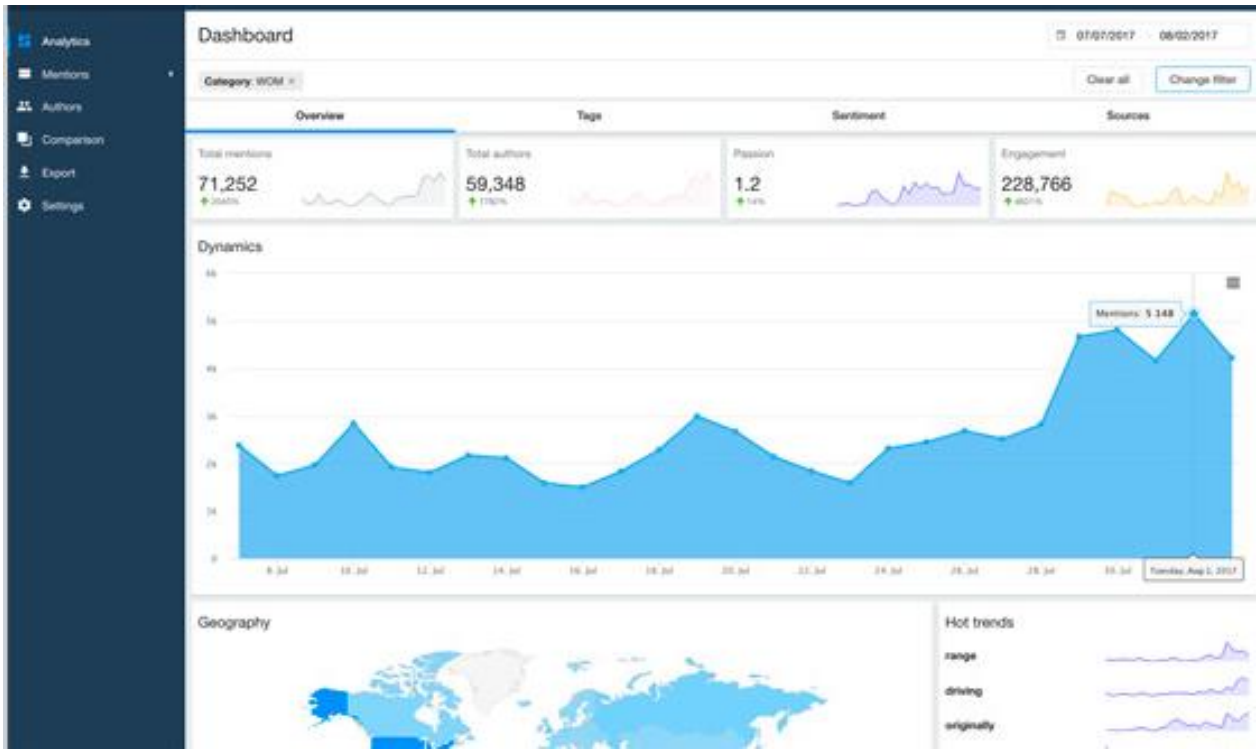

Рисунок 1.11 – Аналитический дэшборд в системе YouScan

YouScan отслеживает все главные социальные социальные сети, такие как Facebook, Twitter, Instagram, а также блоги и форумы, сайты отзывов, онлайн-СМИ и даже мессенджеры, например, Telegram. При этом благодаря использованию технологий компьютерного зрения, система может распознавать помимо текста, еще и графические изображения, учитывая при этом их контекст, тональность и характер самого изображения. Например для поиска информации о бренде, достаточно чтобы на картинке находился логотип этого бренда и эти данные также попадут в сводную статистику.

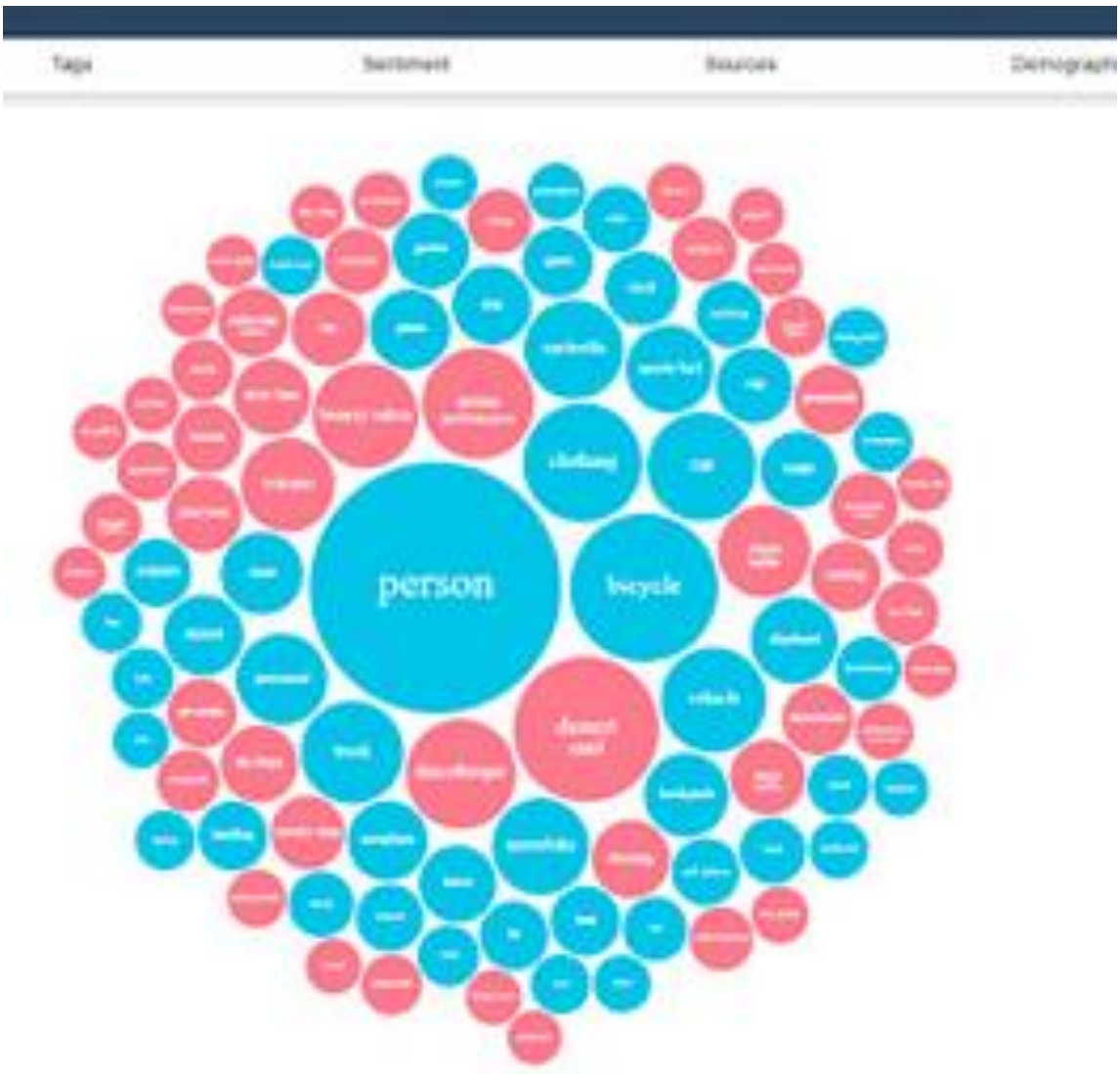

Рисунок 1.12 – Облако тегов после обработки изображения в YouScan

Помимо крупных зарубежных систем, существуют и казахстанские решения по аналитике данных в СМИ и социальных сетях. Их работу можно рассмотреть на примере системы iMAS, которая применялась в проектах связанными с государственными органами, такими как акиматы разных областей и прокуратура [77].

Акимат города Алматы, занимаясь улучшением своей деятельности и повышением информированности о проблемах города и населения применил данную систему для анализа настроения и тематики 21233 сообщений от жителей. А также данная система позволила проанализировать 1700 видеоматериалов и 4816 негативно настроенных комментариев.

Акимат города, занимаясь улучшением своей работы и повышением качества госуслуг, воспользовался услугами системы iMAS. Главная задача, которая стояла перед нами - информирование акимата по всем недочетам по городу, выявление настроения и критики со стороны населения. Результаты

данной аналитики позволяют корректировать работу отдельных служб города и перераспределять ресурсы.

Министерство юстиции РК используя тот же подход проводит аналитику основных тенденций и настроений общественности, а также проводит оценку своих собственных действий в зависимости от их оценки в социальных сетях и характере упоминаний. Так, например, они проанализировали около 10000 сообщений для оценки действий своих подразделений, в различных регионах Казахстана.

Как показывают все эти примеры, возможностей для использования алгоритмов распространения информации в распределенных системах достаточно много. Начиная от корректировки действий коммунальных служб в зависимости от сообщений в Instagram и заканчивая выявлений фейковых новостей во время чрезвычайных ситуаций с помощью лингвистического анализа.

Как итог данной главы, можно сказать что были рассмотрены поэтапно все базовые элементы теории диффузии информации, начиная от понятия связи между двумя узлами в сети, и до паттернов построения сетей. Также было уделено значительное внимание тому как изучение распространения информации среди отдельных индивидов может пересекаться с характером поведения СМИ в такой же ситуации, а также влияние медиа ресурсов на поведение индивидов. Далее на основе исследований в области предотвращения распространения недостоверной информации в крупных распределенных сетях, были проиллюстрированы все подходы в анализе контента, которые будут применены далее в практической части данной работы.

И как финальный этап были рассмотрены варианты применения данных научных исследований в коммерческих продуктах, которые позволяют крупным компаниям отслеживать свою репутацию среди их целевой аудитории, а казахстанским государственным органам корректировать свою работу в зависимости от реакции населения.

Все это говорит, как о высокой востребованности проведения исследований в данной области, так и значительной практической применимости проводимого анализа.

2 Анализ распространения информации в казахстанских СМИ при возникновении крупных инфоповодов

В качестве использования описанных ранее подходов в анализе диффузии информации далее в данной диссертационной работе будет проведен анализ распространения новости среди средств массовой информации об отставке первого президента Республики Казахстан – Назарбаева Н.А.

Этот инфоповод был выбран неслучайно, так как является самым крупным в 2019 году, а, следовательно, по нему возможно собрать необходимый набор данных, достаточный для проведения анализа. Из перечисленных ранее методов будут использоваться подходы, связанные с лингвистическим анализом и расстановкой меток, которые были рассмотрены в части связанной с анализом распространения фейковых новостей.

Уникальность данного подхода по сравнению с рассмотренными в этой главе готовыми коммерческими решениями в том, что результатом этого исследования будет являться воссозданная картина распространения информации в длительной ретроспективе. Когда как все инструменты аналитики такие как Google Alerts, YouScan и другие, направлены на получение данных в реальном времени, и не позволяют выявлять взаимосвязи в таком длительном временном периоде.

Также такая воссозданная картина позволяет определить взаимосвязи между источниками распространения информации, и выявления ключевых звеньев в процессе распространения информации. Как показали рассмотренные ранее исследования, обычно подобная аналитика проводилась в крупных социальных сетях, таких как Facebook и Twitter, но еще ни разу в отношении казахстанских СМИ. А также в отличие от указанных исследований, данный анализ будет содержать данные о временном распределении информации между медиа ресурсами.

Результаты данного исследования позволят выявить влияние одних СМИ на другие, а также определить процент схожести их текстов при публикации материалов по одной и той же тематике. А проведя анализ общей статистики посещаемости ресурсов, возможно сформировать некий рейтинг СМИ, на основе которого могут строиться коммерческие решения для маркетинговых исследований при планировании бюджетов.

2.1 Определение числа участников сети распространения информации в Республике Казахстан

Для определения возможного поведения социальной сети в целом, первым этапом необходимо определить ее максимальные границы. При самом оптимистичном сценарии каждый житель страны будет являться агентом распространения информации. Следовательно, первоначальный фильтр при

определении границ сети диффузии информации является абсолютные значения количества населения по областям.

Основным государственным органом в Казахстане, отвечающим за подсчет числа населения является комитет по статистике Республики Казахстан. В соответствии с его данными общее распределение населения по регионам по данным на 2018 год представляет собой следующий график:

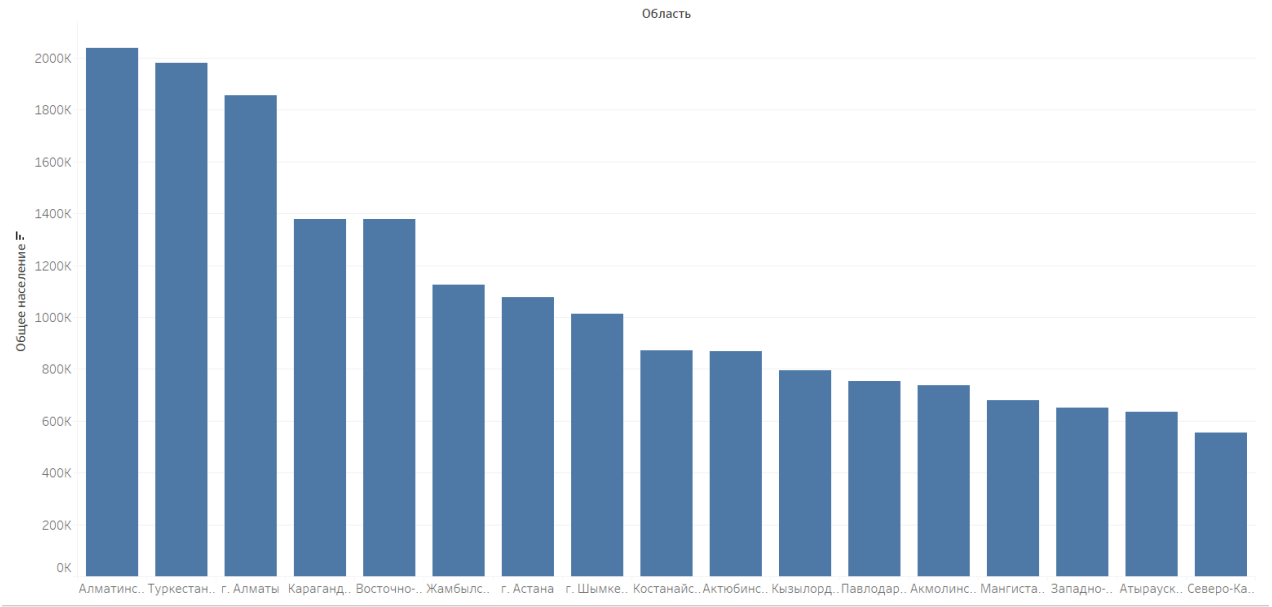

Рисунок 2.1 – Распределение населения по областям

При этом в Республике имеется тенденция на увеличение доли городского населения, а также увеличение численности в целом. Это в свою очередь влияет и на расширение границ максимально возможной сети распространения информации [80].

Прогнозной схемой территориально-пространственного развития страны до 2020 года, влияют на неравномерность прироста. Так, в городах численность населения прирастает быстрее, чем в селах - годовой рост составил 1,7 процента, до 10,43 миллиона человек, и 0,8 процента, до 7,73 миллиона человек, соответственно.

Однако необходимо отметить что население в стране распределено неравномерно, что тоже накладывает некие ограничения при формировании общей сети коммуникации между информационными агентами за счет географической отдаленности друг от друга.

Для определения основных узлов сосредоточения населения статическая информация в виде соотношения Область/Город – Количество населения была наложена на карту Республики Казахстан.

Можно заметить насколько неравномерно распределено население по всей территории РК, а именно: существует значительное сосредоточение население между тремя условными центрами: городами Алматы, Шымкент и Нур-Султан (Астана). Это связано в том числе и с экономическим развитием

этих городов, что в свою очередь влияет и на активность населения участия в общественных активностях и вовлеченность в социокультурные процессы.

При оценке влияния этого фактора на размер и поведение сети распространения информации первый приоритет при анализе первых возникающих потоков информации будет отдан городам Алматы, Шымкент и Нур-Султан.

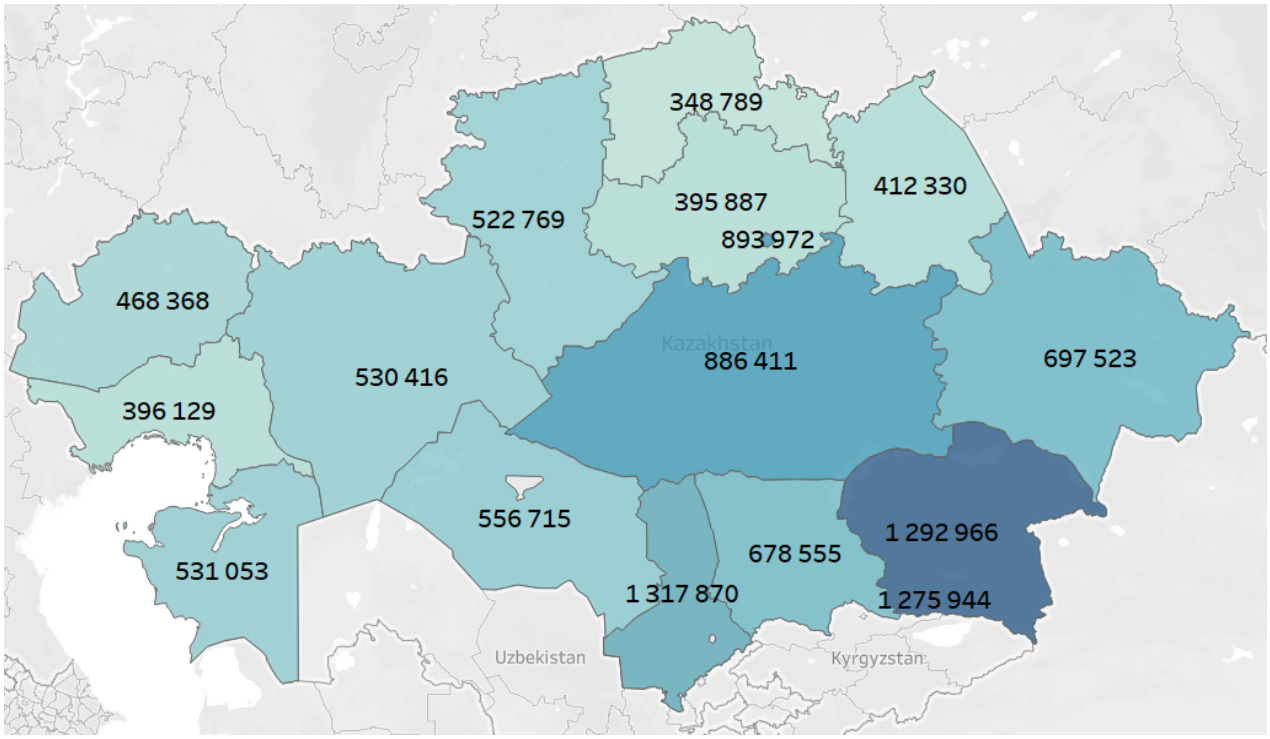

Рисунок 2.2 – Распределение населения по регионам

Однако при оценке размера сети распространения информации необходимо учитывать не только общее количество населения, но и уровень компьютерной грамотности. Даже если в регионе и существует большое сосредоточение агентов, далеко не все смогут действительно поучаствовать в процессе распространения информации. Это обуславливается как уровнем существующей системы образования, так и наличием соответствующего окружения. Так, например, в крупных городах уровень компьютерной грамотности будет выше, просто за счет наличия компьютерных классов в школах.

Для дополнительной фильтрации потенциальных агентов сети распространения информации рассмотрим данные комитета по статистике и просуммируем количество населения, обладающего средними и выше навыками компьютерной грамотности [81].

Рисунок 2.3 – Распределение компьютерно-грамотного населения РК

Как можно увидеть, основная часть участников сети распространения информации находится на юге и юго-востоке страны, а также в крупных городах в центральной части.

Третьим фактором распространения информации является доступ к мобильному и фиксированному интернету.

На текущий момент в Казахстане развивается уже 5G, однако далеко не во всех частях республики имеется доступ к сетям даже третьего поколения. В связи с большим объемом необходимых инвестиций для технологии 5G на первом этапе ее внедрение планируется в городах республиканского значения: Нур-Султан, Алматы и Шымкент. То есть в тех самых городах, в которых как мы выяснили ранее находится самое большое количество компьютерно-грамотного населения. С 2023 года будет производиться уже повсеместное внедрение сетей нового поколения. В результате проводимой работы 97 процентов населения Казахстана до конца 2022 года получат широкополосный доступ к Интернету. Для оставшихся 3 процентов населения будет обеспечена техническая возможность применения спутниковых технологий. Повышение доступности сети Интернет в сельских населенных пунктах является базовым элементом реализации других проектов государственной программы "Цифровой Казахстан", таких как электронное здравоохранение, оказание электронных услуг [82].

Процесс повышения доступности интернета во всех регионах как расширяет саму сеть взаимодействия между агентами, так и ускоряет процессы внутри нее. Тот же самый обмен медиа контентом напрямую зависит от

имеющейся скорости интернета. Далеко не каждый агент будет загружать и распространять контент, если в ходе взаимодействия с ним будут возникать технические трудности.

Соответственно, последний фактор, который является фильтром для количества агентов информационной сети – это наличие технического доступа к сети, для людей, которые имеют навыки взаимодействия с ней.

На конец 2018 года процент обеспеченностью доступом в интернет по данным комитета по статистике составляет 82% [83]. Соответственно текущий размер сети распространения информации представляет собой:

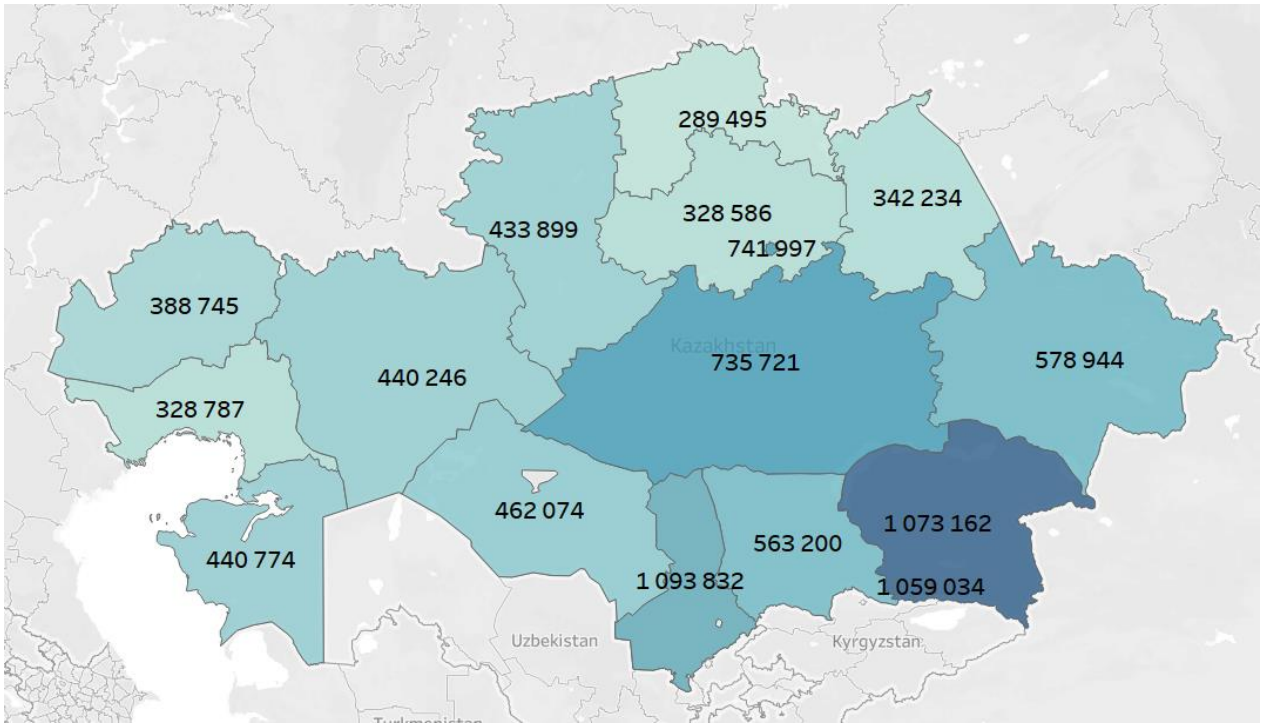

Рисунок 2.4 – Общий размер сети распространения информации

Из этого можно сделать вывод: после определения количества социально активного населения, которое имеет необходимый уровень технической грамотности и доступ к качественному интернет соединению общий размер сети распространения информации составляет 9 миллионов 780 тысяч потенциальных агентов сети.

Важность процесса распространения информации трудно переоценить. В зависимости от характера передаваемых данных ускорение этого процесса может как навредить, в случае дезинформации или вбросов, или же наоборот принести пользу, в случае возникновения чрезвычайной ситуации, требующей немедленных действий со стороны агентов сети.

Но перед тем как анализировать сам процесс распространения информации по сети, необходимо было определить ее границы и уточнить описание потенциальных участников, что и было проделано в этой статье.

В результате удалось определить, что суммарный размер этой сети составляет около 10 миллионов человек. В дальнейшем эта цифра вырастет за

счет повышения уровня компьютерной грамотности населения и процента проникновения широкополосного доступа в интернет во все регионы страны. Сейчас же наиболее активными городами в плане распространения и генерирования информации можно назвать Алматы, Нур-Султан, Шымкент, Туркестан и Караганду а также Алматинскую и Акмолинскую области. Соответственно при анализе вопросов распространения информации стоит сосредоточиться на информационных ресурсах, которые базируются именно в этих городах.

Эти данные в дальнейшем будут использованы для следующего этапа аналитики диффузии информации, а именно: определения инструментов взаимодействия и точек контакта между агентами, анализа характера этого взаимодействия и как результат построение системы анализа данных, которая позволит проводить как ретроспективный анализ распространения информации, так и наоборот определять возможные будущие тренды на ранних стадиях.

В свою очередь распространение информации можно будет рассматривать как взаимодействие разной степени значимости и силы между агентами сети с помощью различных инструментов, такими как СМИ, социальные сети и мессенджеры.

2.2 Аналитика рынка СМИ в Республике Казахстан по ключевым параметрам, влияющим на распространение информации

В данном разделе будет проведено исследование рынка СМИ Республики Казахстан, для выявления ресурсов, которые имеют самый высокий потенциал в процессе распространения информации.

Для этого необходимо проанализировать список возможных метрик применимых к СМИ, сравнить их по этим параметрам, а далее соотнести полученные результаты с ранее рассчитанными данными потенциальных агентов распространения информации в распределенной сети. В качестве первого шага, необходимо дать определение понятию эффективности медиа, а также определить перечень изучаемых метрик.

Так, эффективностью медиа часто называют относительную стоимость размещения на различных площадках, и она может быть измерена медиа-планировщиками и покупателями как цена за тысячу показов, цена за тысячу показов медиаконтента или стоимость привлечения одного покупателя. Эти факторы являются основой, на которой планировщики или покупатели решают, какие средства массовой информации необходимо задействовать в процессе распространения информации.

Тем не менее, это определение не совсем лучший способ анализа эффективности СМИ. Данный подход заставляет планировщиков и покупателей тратить деньги на косвенную стоимость, а не на прогнозируемый результат охвата аудитории. Эффективность чего-либо зависит от его способности достигать желаемых результатов без потерь материалов, времени

и даже энергии. СМИ должны быть проанализированы с точки зрения того, могут ли они обеспечить более высокие показатели осведомленности аудитории по более низкой цене, чем другие варианты. В этом определении заложена перспектива соотношения затрат и выгод, и необходимо определить, возможно ли измерить, сколько людей увидят контент соизмеряя вложенные ресурсы в распространение информации.

При этом будет также ошибочно принимать решения исключительно на основе стоимости аудитории медиа-носителя. Это ошибка в первую очередь связана с тем, что даже если вы задействуете самый дешевый или доступный ресурс распространения информации, и он при этом не работает, то вложенные ресурсы начинают расходоваться неэффективно [84].

Привлечение аудитории часто рассматривается как необходимый шаг к достижению более широких, долгосрочных социальных изменений и, как таковое, часто учитывает краткосрочные критерии оценки, которые используются для оценки результатов работы СМИ. Следует подчеркнуть, что, хотя эти две концепции взаимосвязаны, «вовлечение - это не то же самое, что воздействие».

Вовлеченность - это не то же самое, что воздействие, но они могут быть тесно связаны между собой в объяснительной журналистике. И поэтому, в некоторой степени и в некоторых контекстах, взаимодействие, возможно, лучше всего рассматривать как полезный прокси для воздействия, поэтому контент и измерение его качества продолжают играть заметную роль в области оценки воздействия средств массовой информации.

Некоторая работа в этой области была направлена на разработку так называемых иерархий взаимодействия, в которых различным аспектам взаимодействия назначаются разные веса в соответствии со степенью, в которой они могут способствовать более конкретным формам воздействия. Например, существует следующая иерархическая система метрик.

Таблица 2.1 – Категоризация существующих метрик в зависимости от вовлеченности участников сети

Опыт	**Эмоции и распространение**	**Участие**	**Действие**
Просмотры	Комментарии	Количество контента созданного пользователями	Организация мероприятий
Подписчики	Оценки пользователей	Идентификационные стратегии	Внесение взносов
Лайки (и другие знаки в значении «мне нравится»)	Подписи	Участие в мероприятиях	Волонтерство

	Регистрации на сайте	Ранняя регистрация	

Продолжение таблицы 2.1

	Количество поделившихся в соц сетях	Приглашение других участников	
	Использований хештега	Поиск подобных материалов	
	Пользовательские голосования		
	Посты в блогах и на форумах		

Как следует из этой таблицы, наиболее ценные формы участия в этой конкретной системе оценки воздействия средств массовой информации - это степень, в которой люди принимают участие в различных формах гражданской активности. Она в свою очередь может быть определена по трем основным критериям:

- Способность получать и обрабатывать информацию, имеющую отношение к формированию мнений по гражданским вопросам;
- Способность высказывать и обсуждать мнения и убеждения, связанные с гражданскими вопросами;
- Способность принимать меры в отношении гражданских вопросов.

Исходя из этих трех широких категорий, гражданское участие осуществлялось различными способами, начиная от участия в гражданских мероприятиях и до участия в различных формах политической коммуникации (например, связываться с государственными должностными лицами, выступать на публичном форуме, подписывать петиции или, в последнее время, участвовать в различных формах онлайн-общения, таких как комментирование новостей и блогов, или участие в политической деятельности через социальные сети).

Учитывая диапазон доступных вариантов, также были предприняты попытки классифицировать диапазон онлайн-показателей, которые можно использовать для оценки вовлеченности. Например, классифицировать диапазон метрик в соответствии с четырьмя всеобъемлющими категориями:

- Вовлечение участников (которое включает в себя такие показатели, как количество и частота действий, таких как загрузка и выгрузка контента; установка виджетов и приложений; частота запуска и завершения видео и т. д.);
- Измерение степени неравнодушия к теме (которое включает в себя такие показатели, как объем онлайн-обсуждения, упоминания в социальных сетях, онлайн-настроения и т. д.);

- Защита обсуждаемой темы (которая включает в себя такие метрики, как количество рекомендаций, обзоры продуктов и отзывы, коэффициенты обмена и т. д.);

- Влияние сети (которое включает в себя такие показатели, как количество подписчиков, ретвиты и скорости передачи, количество входящих ссылок и т. д.) [85]

Получившийся результат представляет из себя формулу:

$$\sum(C_i + D_i + R_i + L_i + B_i + F_i + I_i) \quad (2.1)$$

Где:

C = Индекс глубины щелчка (получен из просмотра страниц и событий).

D = индекс продолжительности (полученный из времени, проведенного на сайте).

R = индекс недавности (определяется исходя из скорости, с которой посетители возвращаются на сайт с течением времени).

L = индекс лояльности (определяется на основе уровня долгосрочного взаимодействия посетителей с сайтом).

B = индекс тематики (определяется на основе узнаваемости бренда посетителем сайта или продукта).

F = индекс обратной связи (получен из качественной информации, включая склонность запрашивать дополнительную информацию или предоставлять прямую обратную связь)

I = Индекс взаимодействия (определяется на основе взаимодействия посетителей с контентом или функциями, предназначенными для повышения внимания посетителей к тематике или сайту).

Как должно быть ясно, многие из критериев, используемых в этом подходе, являются производными от более традиционных показателей «подверженности»; и есть сомнения, могут ли критерии, связанные с освещением в аудитории, сами по себе эффективно представлять более сложную конструкцию, такую как вовлечение аудитории. Представляется, что понятие вовлечения предполагает измерения взаимодействия аудитории со СМИ, которые выходят за пределы частоты и продолжительности воздействия.

Сама оценка ценности анализируемой статьи может заключаться в следующем:

- Количество показов (то есть, просмотр сообщений бренда в социальной среде);

- Показы страниц - и личные действия (т. е. взаимодействие с контентом посредством действия, такого как нажатие на фото, видео или ссылку);

- Публичные действия (т. е. распространение контента бренда через платформы социальных сетей через лайки, комментирование и т. д.).

Как результат эти показатели можно взять за основу в процессе сравнения степени влияния СМИ на аудиторию [86].

После того как мы определили критерии оценки СМИ, можно сделать на их основе анализ рынка Республики Казахстан. Так как изучаемое нами

событие, а именно: отставка первого президента РК – Назарбаева Н.А. произошло в 2019 году, следовательно, далее будут рассмотрены данные именно за этот период.

Анализируемые данные представлены в таблице ниже [87].

Таблица 2.2 – Данные по рынку СМИ в Республике Казахстан

Дата	Просмотры	Сессии	Посетители	В среднем online	В среднем активных online
ноя.17	6367235	1913901	1274652	26627	14550
дек.17	6167426	1856033	1229291	25265	13718
янв.18	6737482	2079991	1355299	28178	14796
фев.18	6561403	2103264	1381569	28467	14608
мар.18	5992859	1965553	1270555	26033	12999
апр.18	6067820	2105939	1331303	28782	14815
май.18	6188538	2151196	1391868	29194	14729
июн.18	6354041	2218841	1456989	29736	14839
июл.18	9313206	2908525	1754852	38967	21178
авг.18	8494825	2715827	1648612	36343	19596
сен.18	8675257	2869495	1794943	37924	19851
окт.18	8536139	2864720	1810063	37847	19466
ноя.18	8179051	2782084	1780567	36609	18621
дек.18	7433385	2586153	1663352	33636	16900
янв.19	8409214	2878053	1812587	37763	19406
фев.19	9322453	3144746	1957871	41502	21595
мар.19	7983258	2892843	1837344	37308	18611
апр.19	8432540	3089469	1994261	39887	19619
май.19	8830247	3205218	2092646	41005	20231
июн.19	10272555	3498765	2177390	45284	23609
июл.19	9321553	3290832	2066320	42714	22134
авг.19	8258300	3059773	1921547	39505	19807
сен.19	9739643	3622692	2268085	46839	23458
окт.19	7208445	2864625	1861738	37053	17816
ноя.19	6288059	2679654	1786756	34288	15629
дек.19	6294031	2660470	1753024	33973	15578
янв.20	7483017	3101534	1978111	39735	18462
фев.20	10121833	3201301	2066523	41196	23548
мар.20	16956552	4903635	2882452	62797	39108
апр.20	20854052	6338301	3553199	80827	48630
май.20	15416761	4738679	2809597	59758	36228

На основе этих данных возможно отследить динамику изменения трех ключевых параметров СМИ: количество просмотров, количество пользователей и количество сессий, а также выявить насколько они между собой коррелируют.

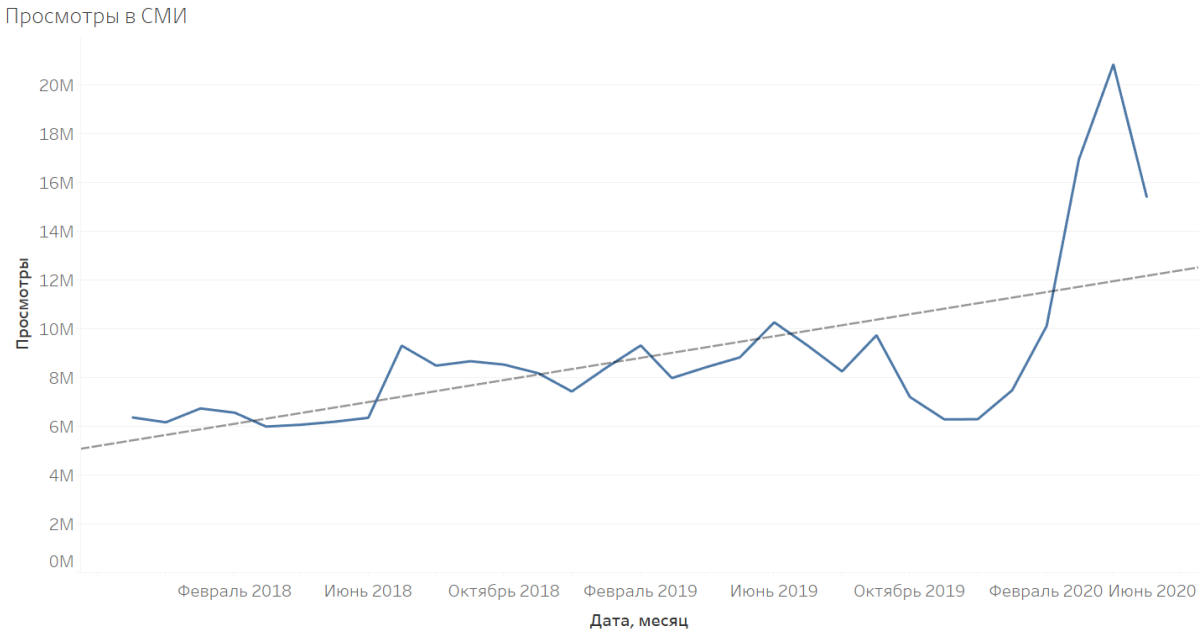

Рисунок 2.5 – Количество просмотров в казахстанских цифровых СМИ

Количество просмотров и посетителей меняется в зависимости от крупных инфоповодов, которые есть в СМИ. Так, например, можно увидеть три всплеска: в июле 2018 года, марте 2019 года, и в марте 2020 года. Наш изучаемый инфоповод как раз и произошел в марте 2019 года.

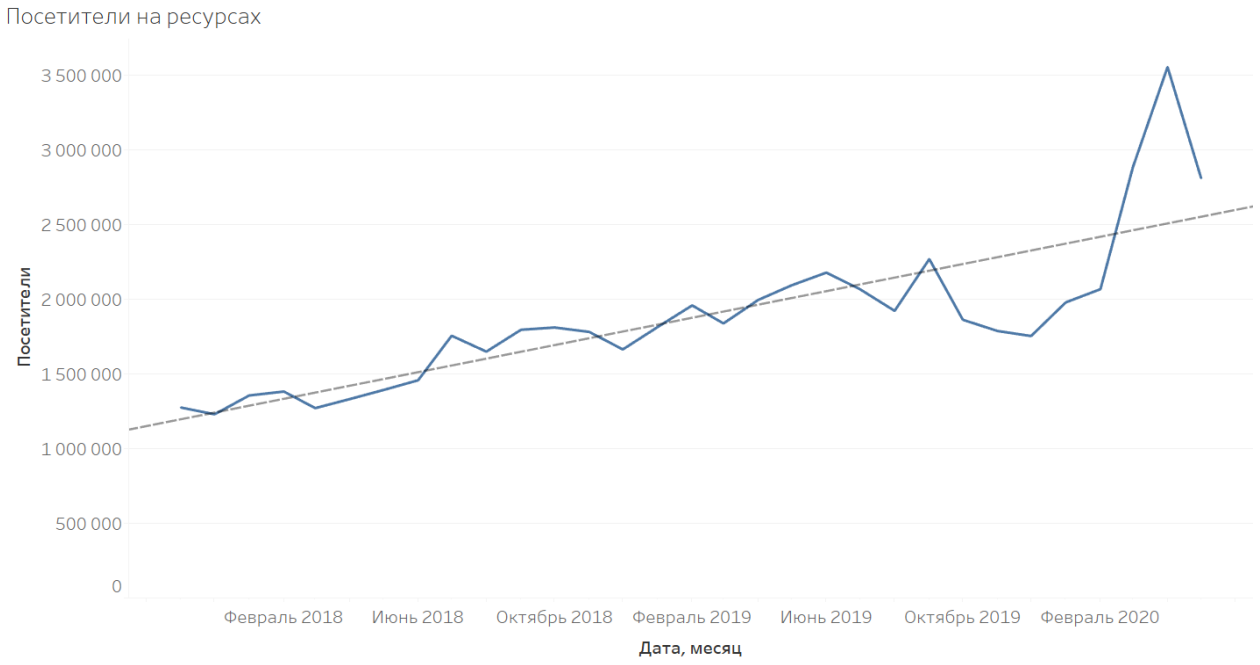

Рисунок 2.6 – Количество посетителей в казахстанских цифровых СМИ

При этом общий тренд на количество посетителей в СМИ стабильно растет уже как минимум два года. Однако в марте 2020 года произошел выброс связанный с коронавирусом, однако уже сейчас можно увидеть что происходит корректировка обратно к линии тренда

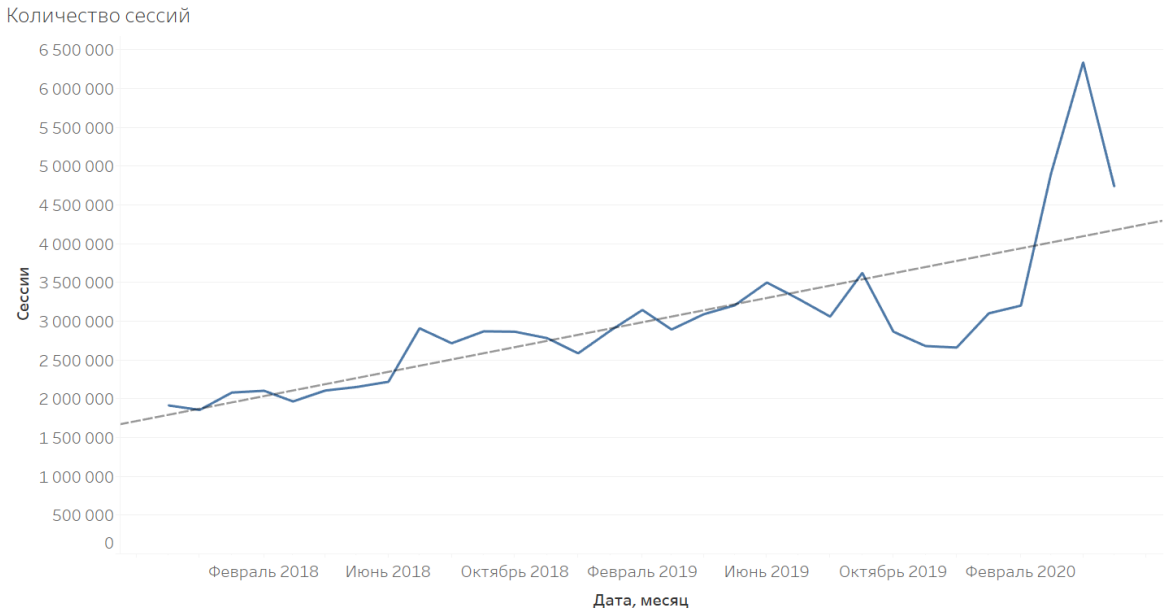

Рисунок 2.7 – Количество сессий в казахстанских цифровых СМИ

Закономерно отметить, что с ростом количества пользователей растет и количество сессий – заходов пользователей на ресурс.

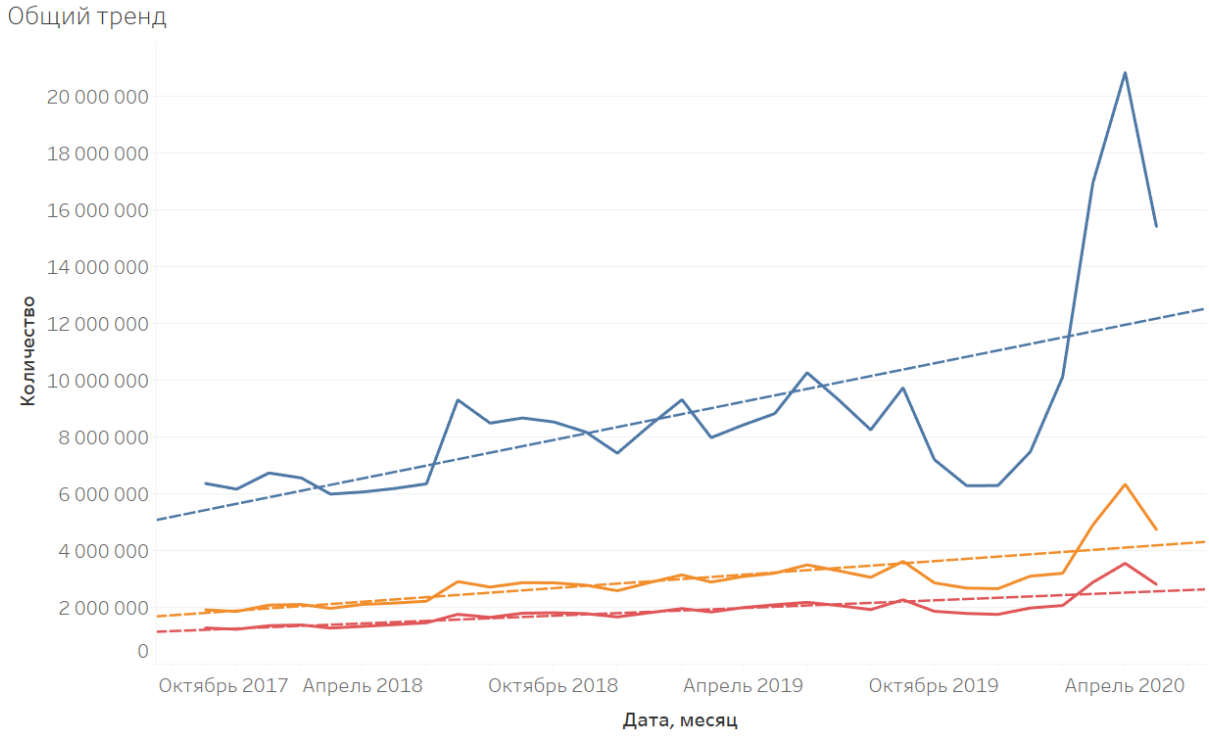

Рисунок 2.8 – Общий тренд трех основных метрик в казахстанских цифровых средств массовой информации

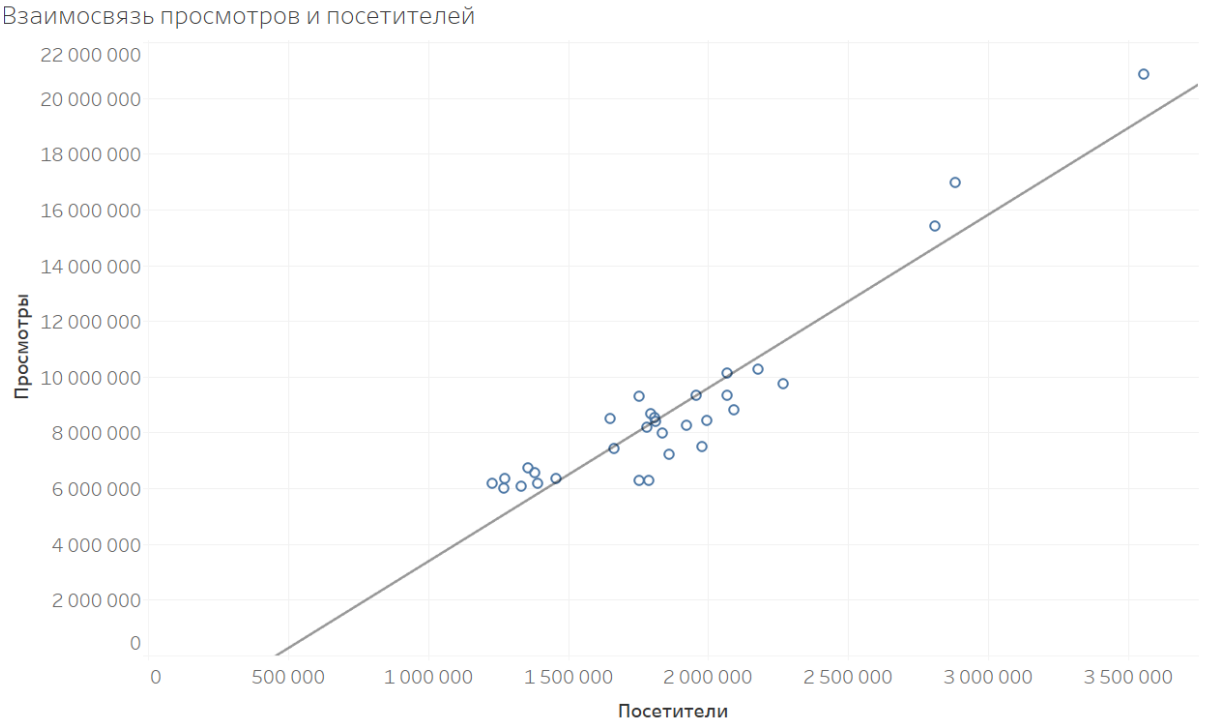

Рисунок 2.9 – Корреляция между количеством посетителей и количеством просмотров

При этом важно проанализировать насколько влияет рост количества пользователей на количество просмотров, и проверить верна ли гипотеза о том, что количество пользователей не выросло, а выросла активность уже существующей базы.

Для того чтобы провести эту оценку, можно рассчитать коэффициент Пирсона. Модель может быть значимой при p <= 0,05.

Таблица 2.3 – Статистические параметры анализа взаимосвязи посетителей и просмотров

Параметр	**Значение**
Количество смоделированных наблюдений:	31
Модели степеней свободы:	2
Остаточные степени свободы (DF):	29
Коэффициент детерминации	0,888879
p-значение (значимость):	< 0,0001

Коэффициент корреляции Пирсона при этом будет равняться квадратному корню из коэффициента детерминации, а, следовательно, будет равен: 0,94.

А значит гипотеза о том, что рост количества просмотров произошел не из-за роста количества пользователей не подтверждается, так есть большая взаимосвязь между этими параметрами.

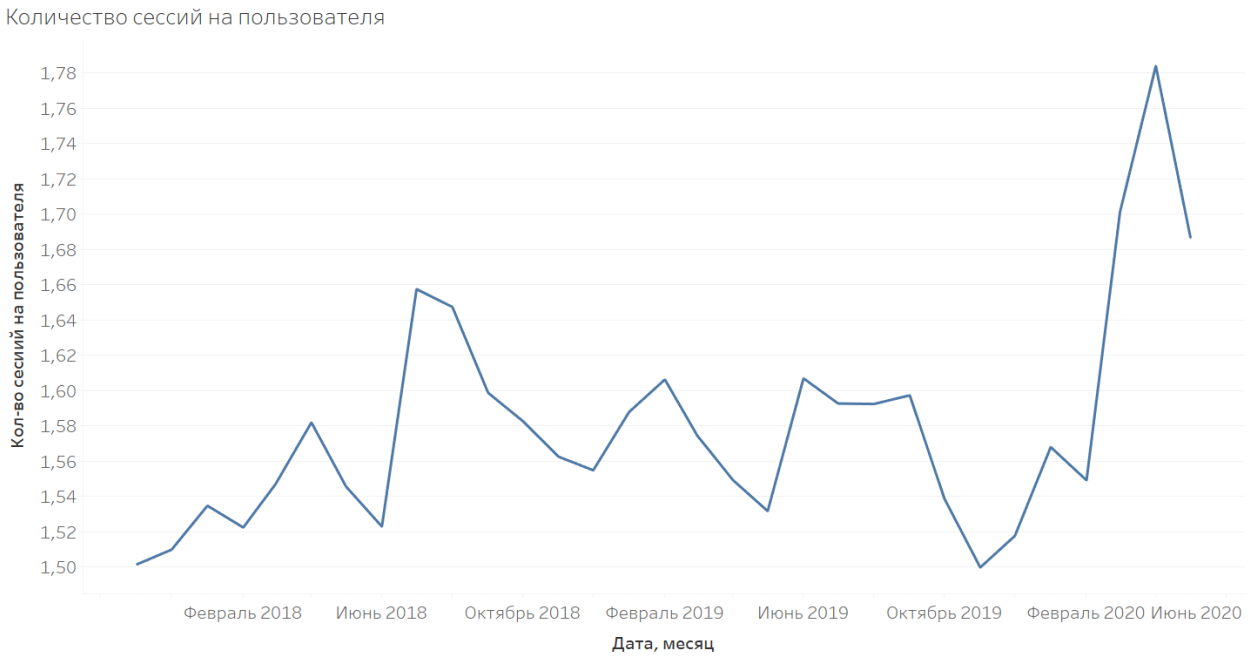

Рисунок 2.10 – Количество сессий на одного посетителя

Динамика количества сессий на пользователя, показывает, что большие инфоповоды сподвигают пользователей в течение некоторого времени чаще заходить на новостные ресурсы, однако пик достигается за довольно короткое время, а дальше идет возврат к среднему значению.

Рисунок 2.11 – Количество просмотров на одного пользователя

При этом значительные инфоповоды влияют еще не только на то как часто будет заходить пользователь на ресурс, но и как много статей он при этом просмотрит.

Как результат, можно сказать что все три параметра СМИ: просмотры, сессии и количество пользователей – тесно между собой взаимосвязаны. А также можно сделать вывод что распространение информации стимулирует то как много людей посмотрели новость (посетители), как много контента они поглотили (просмотры) и как часто они это делают (сессии).

2.3 Анализ списка популярных СМИ по ключевым параметрам для распространения информации

Ранее были определены основные параметры, по которым можно оценить результативность работы СМИ. Ими стали: количество просмотров, количество посетителей и количество пользовательских сессий. На основе этих параметров можно определить результативность работы 25 лучших СМИ в Республике Казахстан.

Для этого на основе статистических данных можно проранжировать СМИ на основе количества новостей, переходов по ним и общей доле от медиа рынка [88]. Получившаяся картина представляет из себя матрицу, где внутри содержатся места в рейтинге, а также периоды времени и название издательств. Так как изучаемое нами событие произошло в 2019 году, то и датасет сформирован тоже за 2019 год. Более подробные данные в разрезе каждого СМИ можно увидеть в приложении А.

Таблица 2.4 – Распределение мест среди 25 самых результативных СМИ в РК

Названия изданий	янв	фев	мар	апр	май	июн	июл	авг	сен	окт	ноя	дек
aktobegazeti	25			22								
alau	2	2	3	5	5	6	4	7	6	8	9	9
altyn-orda	15	11	9	10	16	21					24	
baigenews										15	6	10
baribar	21	20	21	21	25	23		23			16	24
bnews	8	9	10	6	4	4	5	2	23			
caravan								5	4	3	7	7
dixinews		25	24									
dknews	10	13	11	11	13	14	10	11	14	14	17	17
elorda	17	14	13	14	11	12	14	14	20	24		
forbes	19	18	19	17	15	17	19	20	16	5	3	4
ia-centr	16	17	16	19	19	18	15	19	17	19	20	18
inbusiness							23	16				
informburo	6	4	5	4	6	5	3	8	7	9	10	5
kapital	24											
kapshagai	23	24										

kaz.tengrinews	13	12	15	16	17	15	11	13	10	12	14	13
knews									25			

Продолжение таблицы 2.4

kp		23	23	25			21	25	22	23	22	23
kstnews	3	6	8	8	9	7	8	9	9	11	12	11
lada	4	3	2	3	3	3	2	3	2	4	2	2
mgorod	11	10	7	9	7	8	9	10	8	6	5	6
mix.tn	20	21	20	20	23	22	22	24	21	22	23	22
mtrk							24			25	25	
ng	12	15	14	13	12	11	13	15	11	16	15	16
online.zakon				24	24	24			24			25
qostanay	7	7	6	7	8		7	6	5	7	8	8
ru.sputniknews	18	19	18	18	18	19	20	22	18	18	18	20
shakhty						20						
smirnovo			22			25						
tengrinews	1	1	1	1	1	1	1	1	1	1	1	1
timeskz		22			22							
today						13	12	17	12	13	13	12
toppress					20	10	17	21	19	20	21	21
tumba	14	16	17	15	14	16	18	18	15	17	19	19
uralskweek	9	8	12	12	10	9	16	12	13	10	11	14
vecher	22						25					
voxpopuli			25	23	21							
vrk												15
zakon	5	5	4	2	2	2	6	4	3	2	4	3
zirki										21		

По таблице выше можно сделать выводы насчет постоянства нахождения СМИ в рейтинге, а, следовательно, это свидетельствует о систематичности работы и нивелирует влияние конкретных новостей на общую статистику.

Таблица 2.5 – Частота появлений СМИ в рейтинге

Названия строк	**Количество появлений в рейтинге**
aktobegazeti	2
alau	12
altyn-orda	7
baigenews	3
baribar	9
bnews	9
caravan	5
dixinews	2

dknews	12
elorda	10
forbes	12

Продолжение таблицы 2.5

ia-centr	12
inbusiness	2
informburo	12
kapital	1
kapshagai	2
kaz.tengrinews	12
knews	1
kp	9
kstnews	12
lada	12
mgorod	12
mix.tn	12
mtrk	3
ng	12
online.zakon	5
qostanay	11
ru.sputniknews	12
shakhty	1
smirnovo	2
tengrinews	12
timeskz	2
today	7
toppress	8
tumba	12
uralskweek	12
vecher	2
voxpopuli	3
vrk	1
zakon	12
zirki	1

Как видно из получившейся статистики существует целый ряд СМИ которые появляются в списке каждый месяц, однако места которые они занимают тоже оказывают значительное влияние на принятие решения о влиятельности данного СМИ. Именно для этого можно просуммировать общие занятые места и разделить на количество значений, и чем меньше будет это значение, тем более востребованным может считаться СМИ.

Однако первичным все же остается параметр связанный с количеством упоминаний в рейтинге, так как он формировался на основе ключевых

параметров, таких как: количество просмотров каждой новости, количество посетителей, количество сессий.

Все эти факторы позволяют сформировать общий рейтинг СМИ в зависимости от того насколько они имеют потенциал в плане распространения информации.

Таблица 2.6 – Список СМИ, упорядоченный по ключевым параметрам, рейтингу, количеству упоминаний в рейтинге

Номер	**Издание**	**Количество упоминаний**	**Сумма количества мест**
1	tengrinews	12	12
2	lada	12	33
3	zakon	12	42
4	alau	12	66
5	informburo	12	72
6	mgorod	12	96
7	kstnews	12	101
8	uralskweek	12	136
9	dknews	12	155
10	kaz.tengrinews	12	161
11	ng	12	163
12	forbes	12	172
13	tumba	12	198
14	ia-centr	12	213
15	ru.sputniknews	12	226
16	mix.tn	12	260
17	qostanay	11	76
18	elorda	10	153
19	bnews	9	71
20	baribar	9	194
21	kp	9	207
22	toppress	8	149
23	today	7	92
24	altyn-orda	7	106
25	caravan	5	26
26	online.zakon	5	121
27	baigenews	3	31
28	voxpopuli	3	69

29	mtrk	3	74
30	inbusiness	2	39

Продолжение таблицы 2.6

31	timeskz	2	44
32	aktobegazeti	2	47
33	kapshagai	2	47
34	smirnovo	2	47
35	vecher	2	47
36	dixinews	2	49
37	vrk	1	15
38	shakhty	1	20
39	zirki	1	21
40	kapital	1	24
41	knews	1	25

Как результат, мы получили список из 41 СМИ, которые хотя бы однажды попадали в список самых востребованных в 2019 году. Однако для дальнейшего исследования будут задействованы только первые 25 ввиду их большей статистической значимости.

2.4 Распределение СМИ и сопоставление их с аудиторией внутри регионов РК

Как ранее было выявлено, в Казахстане около 10 миллионов потенциальных агентов распространения информации в распределенной сети. Однако они рассредоточены по территории РК неравномерно. При этом уровень взаимодействия со СМИ у них тоже разный. Можно явно проследить взаимосвязь между количеством средств массовой информации и уровнем культурной и гражданской активности в регионе.

Следовательно, необходимо соотнести сформированный ранее список 25 самых востребованных СМИ с теми регионами в которых они находятся.

Для это можно воспользоваться сервисами whois чтобы понять в каком городе зарегистрирован доменный адрес и где находится юридический адрес этих организаций. С учетом проведения исследований с помощью указанного сервиса, удалось выяснить местоположение каждого из указанных в списке ресурсов. При этом 23 из них находятся на территории Республики Казахстан а еще 2 имеют юридический адрес в Российской Федерации, однако часть их контента также указывает на действие на территории РК.

При этом с помощью такого исследования можно построить карту наибольшей активности самых востребованных СМИ в 2019 году в разрезе регионов Республики Казахстан.

Сформированные списки рейтинга СМИ позволяют нам сосредоточиться исключительно на основных новостных ресурсах, которые формируют повестку дня, однако не стоит пренебрегать и региональными небольшим новостными агентствами, которые могут освещать происходящие события местного масштаба. И как правило оказывающих значительное влияние на распространение информации. Однако изучаемое нами событие было республиканского масштаба, и поэтому нам интересны именно крупные СМИ.

Таблица 2.7 – Распределение СМИ между регионами

Номер	**Издание**	**Город местонахождения**
1	tengrinews	Алматы
2	lada	Актау
3	zakon	Алматы
4	alau	Костанай
5	informburo	Алматы
6	mgorod	Уральск
7	kstnews	Костанай
8	uralskweek	Уральск
9	dknews	Алматы
10	kaz.tengrinews	Алматы
11	ng	Костанай
12	forbes	Алматы
13	tumba	Актау
14	ia-centr	Москва
15	ru.sputniknews	Москва
16	mix.tn	Алматы
17	qostanay	Костанай
18	elorda	Нур-Султан
19	bnews	Нур-Султан
20	baribar	Караганда
21	kp	Алматы
22	toppress	Нур-Султан
23	today	Алматы
24	altyn-orda	Алматы
25	caravan	Алматы

В результате нам удалось сформировать список информационных ресурсов, который распределен в зависимости от места регистрации и

юридического адреса. Это не означает что указанные ресурсы не имеют влияния в других регионах, и даже совсем наоборот, однако можно с уверенностью сказать, что наличие данных СМИ в указанных регионах, оказывает влияние еще и на распространение информации местного назначения, так как площадка одна и та же.

Далее мы можем провести сравнение, между тем как распределены СМИ между регионами и тем какое количество населения проживает в данной местности.

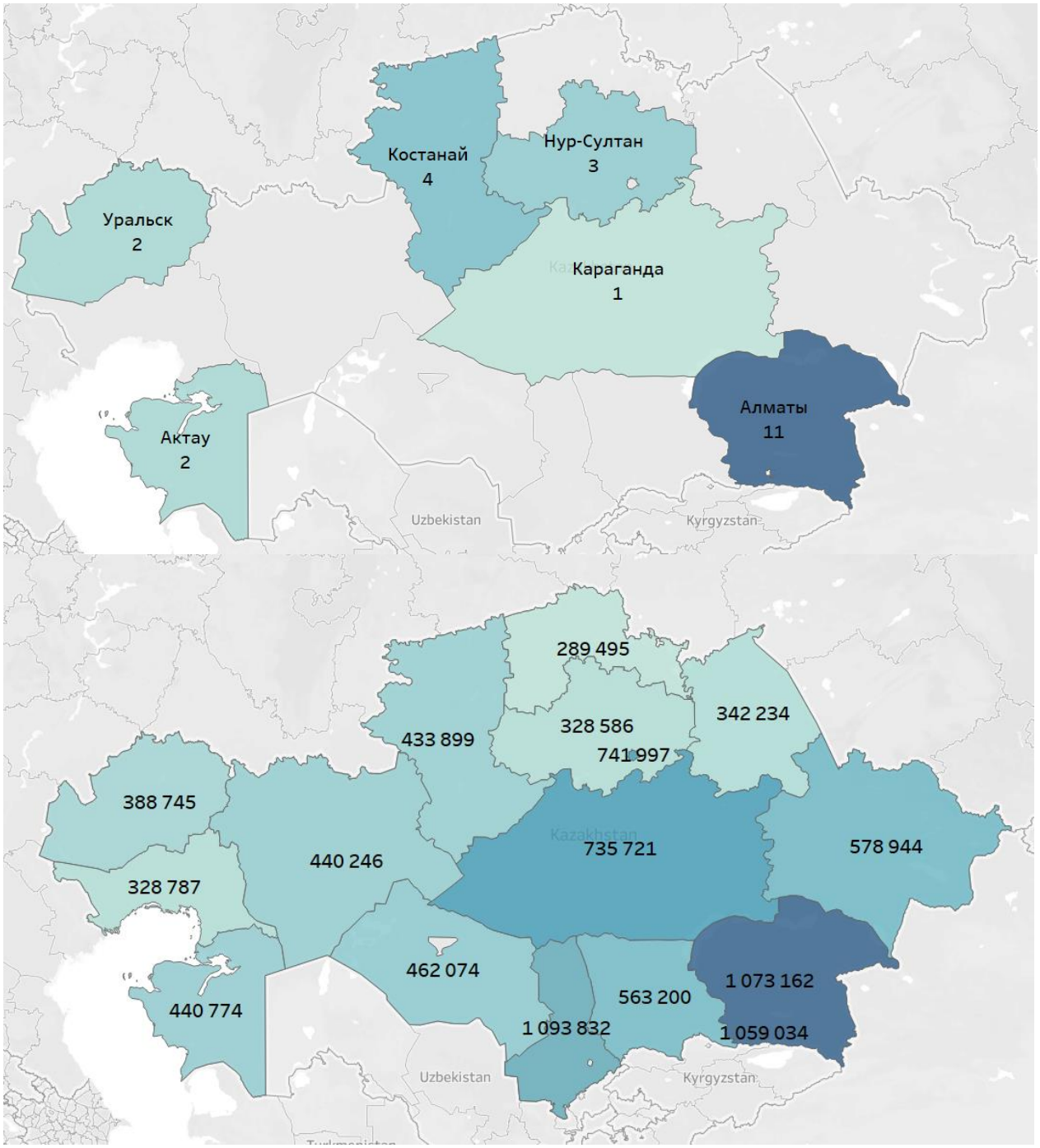

Рисунок 2.12 – Распределение количества СМИ по РК (сверху) и их потенциальная аудитория в самом регионе (снизу)

Как видно на рисунке 2.12 существуют пересечения между количеством активного проживающего населения в регионе а также количеством СМИ которые находятся там. При этом как можно увидеть на рисунке существуют несколько крупных объединений средств массовой информации в Алматы, Нур-Султане, Костанае, Караганде, Актау и Уральске. При количество агентов распространения информации в этих регионах составляет 55% от общего числа.

Далее мы сосредоточимся на изучении новостей, которые были посвящены тематике отставки первого президента Республики Казахстан и которые при этом были размещены на этих ресурсах. Следующими шагами является выяснение степени схожести между статьями и разложение их на временную шкалу.

2.5 Поиск данных в СМИ касающихся анализируемого события

В предыдущем разделе был сформирован список из 25 самых востребованных СМИ в Республике Казахстан, с учетом параметров количества просмотров, посетителей, сессий, количества новостей и переходов по ним, а также с учетом количества населения в регионе потенциально имеющего доступ к этому информационному ресурсу.

Теперь необходимо определить на каких ресурсах и в какое время были размещены статьи по изучаемой нами тематике: прекращение полномочий первого президента Республики Казахстан Назарбаева Н.А. В ходе исследования и поиска по ключевым словам, по определенному веб-ресурс с помощью операторов «site», «before» и «after» удалось выяснить что статьи были размещены на 19 из 25 ресурсов. На остальных информация об этом события появилась в более позднее время и в опосредованном виде.

Причем первоначальным источником является ресурс который не участвует в рейтинге, так как является официальным государственным ресурсом по информированию населения: akorda.kz . На данном ресурсе размещаются официальные тексты заявлений правительства и в нашем случае был размещен полный текст обращения президента. В дальнейшем все источники в том или ином виде будут ссылаться на текст этого документа, и поэтому при анализе статей не учитываются прямые цитаты официального заявления, а только уникальное текстовое описание.

Такое отделение уникальной составляющей текста каждой статьи позволяет сделать анализ сходства статей между собой более прозрачным, и он не будет подвержен погрешности, связанной с обязательной общей составляющей, которая есть на каждом ресурсе.

После того как был сформирован необходимый датасет состоящий из названия ресурса, даты публикации, уникальной части статьи и номера в рейтинге, статьи были отсортированы от самой ранней до самой поздней, не принимая во внимание первоисточник.

Изначальному тексту обращения было уделено отдельное внимание и удалось выяснить, что первоначально статья была размещена в 18:01, когда

попала в обработку web.archive.org., однако в открытом доступе она появилась одновременно с началом обращения в 19:00. Данные о дате публикации удалось получить с помощью сервиса [89], который используя информацию о первых сохраненных копиях веб-страницы на различных поисковых платформах, выявляет самое раннее появление и выводит эту информацию в формате JSON.


```
"sources": {
  "web.archive.org": {
    "uri-m": "https://web.archive.org/web/20190319180135/http://www.
    "memento-datetime": "2019-03-19T18:01:35",
    "memento-pubdate": "",
    "earliest": "2019-03-19T18:01:35"
  },
  "bing.com": {
    "earliest": ""
  },
  "bitly.com": {
    "earliest": "2019-03-19T14:09:21"
  },
  "google.com": {
    "earliest": "2019-03-19T23:59:59"
  },
```


Рисунок 2.13 – Данные в формате JSON о появлении статьи на государственном ресурсе akorda.kz

Примечательно, что информация о прекращении полномочий на ресурсе zakon.kz появилась всего через минуту после появления сообщения сайте Акорды. Это может говорить об автоматическом характере обработки новостей с государственного ресурса.


```
"sources": {
  "web.archive.org": {
    "uri-m": "https://web.archive.org/web/20190320130033/https://www.
    "memento-datetime": "2019-03-20T13:00:33",
    "memento-pubdate": "2019-03-19T19:01:22",
    "earliest": "2019-03-19T19:01:22"
  },
  "bing.com": {
    "earliest": ""
  },
  "bitly.com": {
    "earliest": "2019-03-19T13:37:17"
  },
  "google.com": {
    "earliest": "2019-03-19T23:59:59"
  },
  "last-modified": {
    "earliest": ""
  },
  "pubdate": {
    "earliest": "2019-03-19T19:01:22"
  },
```

Рисунок 2.14 – Данные в формате JSON о появлении статьи на государственном ресурсе zakon.kz

Однако информационный портал zakon.kz стал только первым агентом распространения информации. Как видно из таблицы 2.8, следующие новости начали появляться уже через 3 минуты после размещения на первом ресурсе. Стоит заметить при этом что подавляющее большинство материалов статей были размещены еще до момента окончания обращения президента в 19:26. Более поздние статьи уже включали в себя пост-аналитику текста обращения, полученную от экспертов.

Таблица 2.8 – Информация по дате публикации новостей в СМИ

Номер в рейтинге	**Ресурс**	**Дата публикации**
3	zakon	19.03.2019,19:01
19	bnews	19.03.2019,19:04
5	informburo	19.03.2019,19:04
23	today	19.03.2019,19:07
7	kstnews	19.03.2019,19:07
1	tengrinews	19.03.2019,19:08
2	lada.kz	19.03.2019,19:08
18	elorda	19.03.2019,19:08
16	mix.tn	19.03.2019,19:08
15	ru.sputniknews	19.03.2019,19:09
21	kp	19.03.2019,19:09
4	alau	19.03.2019,19:11
20	baribar	19.03.2019,19:15
25	caravan	19.03.2019,19:15
12	forbes	19.03.2019,19:20
13	tumba	19.03.2019,19:37
9	dknews	19.03.2019,20:00
14	ia-centr	19.03.2019,21:32
6	mgorod	20.03.2019,00:01

Однако необходимо отметить, что нет явной корреляции между нахождением в рейтинге 25 самых востребованных СМИ и датой публикации. Так хотя и первым ресурсом стал, который находится наверху списка, сразу за ним следуют источники под номерами 19 и 23. Но в нашем исследовании учитывается не только скорость появления информации, но и способность ресурсов распространять эту информацию дальше.

Иными словами, даже если информация появилась в каком-то источнике раньше остальных, далеко не факт, что распространение будет происходить активно, если аудитория данного ресурса невелика. Это связано, как и с

доверием к крупным авторитетным источникам, так и с возможностью освещения не только на самом портале, но и в социальных сетях, в которых находится также большая аудитория средств массовой информации.

2.6 Определение степени схожести между новостными статьями по одной тематике на разных ресурсах

Для того чтобы с уверенностью сказать, что распространение информации происходило из анализируемого источника, необходимо провести исследование на наличие схожих наборов текстовых данных. Для этого как раз и используются различные алгоритмы по выявлению плагиата и определения дубликатов.

Перед тем как обрабатывать, имеющиеся у нас данные, необходимо провести их предварительную подготовку, а именно:

- удалить все знаки препинания;
- преобразовать все буквы в строчные;
- преобразовать дублирующие пробелы в один, а также удалить пробелы в начале и конце строк;
- разделить текстовую строку на отдельные слова.

Первое понятие, которое используется в алгоритме сравнения это шинглинг, распространенная техника представления документов в виде наборов. Анализируя документ, можно сказать, что его k-шингл является всей возможной последовательной подстрокой длины k, найденной в нем. Еще одна вещь, которую стоит отметить, состоит в том, что набор k-шинглового документа должен состоять только из уникальных k-шинглов.

Шингл — это кусочек текста, размером в несколько слов. Шинглы идут внахлёст друг на друга, поэтому они и называются таким образом (англ., shingles — чешуйки, черепички) [90].

Если несколько шинглов у текстов сходятся, то можно считать, что эти тексты пересекаются между собой. При этом существует прямая пропорциональность между количеством одинаковых шинглов и тем как много одинакового текста в анализируемых документах. Индекс при этом будет осуществлять поиск документов, которые будут обладать максимальным количеством пересечений с анализируемым текстом.

Подобный метод анализа текстов активно применяется в различных системах проверки на плагиат и заимствования. Так как задача у нас схожая, то и алгоритм достаточно точно может указать сходство между двумя и более статьями на информационных ресурсах. Листинг алгоритма приведен в приложении Б.

Проведем последовательный анализ с переключением количества слов в одном шингле от 1 до 4.

Так как невозможно написать статью по такой узкой тематике, не используя схожие слова и конструкции поэтому в процессе построения сети

распространения информации мы будем учитывать только крупные повторяющиеся конструкции.

	zakon	bnews	informburo	today	kstnews	tengrinews	lada.kz	elorda	mix.tn	ru.sputnik	kp	alau	baribar	caravan	forbes	tumba	dknews	ia-centr	mgorod
zakon		12,41%	11,71%	11,85%	77,42%	17,83%	28,43%	20,43%	18,68%	9,17%	12,24%	10,81%	74,60%	23,81%	11,30%	56,10%	23,60%	9,90%	12,17%
bnews			9,46%	15,24%	12,98%	13,78%	10,46%	14,50%	16,80%	11,35%	13,85%	12,15%	12,12%	10,86%	11,41%	16,08%	10,45%	9,56%	12,08%
informburo				12,77%	10,28%	12,87%	11,20%	10,09%	10,48%	19,81%	10,28%	18,42%	11,32%	13,10%	10,57%	15,25%	15,69%	14,42%	28,04%
today					12,40%	32,53%	17,73%	36,11%	33,64%	10,00%	19,83%	13,59%	13,28%	13,02%	28,35%	21,64%	17,89%	8,15%	14,69%
kstnews						14,74%	30,21%	21,84%	20,00%	8,65%	13,04%	11,76%	82,46%	25,00%	11,93%	60,53%	25,30%	10,53%	12,84%
tengrinews							16,28%	30,22%	39,37%	11,24%	14,74%	11,85%	14,74%	13,07%	15,88%	21,60%	14,01%	9,09%	14,45%
lada.kz								19,81%	18,27%	7,20%	20,19%	10,23%	28,87%	24,82%	15,32%	38,83%	26,26%	10,53%	15,20%
elorda									60,00%	13,86%	15,22%	20,00%	21,84%	15,15%	20,39%	22,77%	15,22%	8,08%	14,68%
mix.tn										13,27%	18,60%	23,33%	20,00%	13,85%	21,21%	25,00%	13,33%	10,75%	16,35%
ru.sputniknews											24,18%	25,00%	8,65%	8,16%	12,93%	12,93%	10,78%	9,62%	18,92%
kp												26,67%	11,83%	12,78%	19,61%	28,42%	15,56%	11,70%	24,24%
alau													13,43%	8,93%	13,25%	16,05%	13,43%	18,46%	28,38%
baribar														23,97%	11,93%	58,44%	25,30%	10,53%	13,89%
caravan															12,75%	25,37%	19,05%	11,03%	20,71%
forbes																18,64%	28,42%	8,85%	18,49%
tumba																	27,08%	13,89%	20,51%
dknews																		9,38%	16,04%
ia-centr																			15,89%

Рисунок 2.15 – Анализ текстов статей с различных информационных порталов на сходство при 1 слове в шингле

	zakon	bnews	informburo	today	kstnews	tengrinews	lada,kz	elorda	mix,tn	ru,sputn	kp	alau	baribar	caravan	forbes	tumba	dknews	ia-centr	mgorod
zakon		4,05%	2,68%	2,27%	68,92%	5,74%	21,49%	5,83%	6,96%	0,75%	2,46%	2,11%	65,33%	16,56%	2,04%	50,52%	17,14%	1,67%	2,72%
bnews			3,14%	3,23%	4,32%	3,92%	2,69%	3,01%	5,03%	3,49%	3,68%	4,44%	3,70%	3,65%	4,30%	5,56%	3,09%	1,84%	4,28%
informburo				5,35%	2,90%	5,31%	3,80%	2,86%	5,26%	4,14%	2,90%	11,76%	2,92%	5,26%	3,09%	5,84%	5,26%	8,59%	14,29%
today					2,42%	13,25%	3,80%	23,02%	14,38%	2,30%	4,97%	6,82%	3,07%	3,65%	14,12%	6,15%	5,03%	2,47%	5,98%
kstnews						5,00%	22,52%	6,42%	7,69%	0,82%	2,70%	2,38%	71,21%	17,81%	2,96%	55,17%	17,89%	1,83%	2,94%
tengrinews							4,04%	15,85%	35,06%	2,34%	3,45%	2,82%	4,50%	3,88%	5,38%	7,44%	4,52%	1,47%	4,42%
lada,kz								4,55%	5,51%	0,69%	3,82%	1,89%	23,85%	16,17%	1,90%	29,75%	18,58%	2,31%	3,18%
elorda									35,71%	2,46%	3,57%	3,53%	5,50%	6,10%	9,30%	7,03%	4,59%	0,89%	4,41%
mix,tn										3,42%	4,67%	5,00%	6,73%	6,25%	7,87%	9,92%	5,77%	2,83%	5,34%
ru,sputniknews											12,84%	11,76%	0,83%	1,12%	3,50%	2,86%	1,68%	2,56%	8,76%
kp												13,16%	1,80%	2,38%	5,30%	6,30%	2,75%	3,74%	11,11%
alau													3,66%	2,13%	4,72%	5,94%	1,20%	15,28%	20,43%
baribar														17,93%	1,47%	50,56%	19,35%	1,85%	3,73%
caravan															3,14%	15,57%	11,11%	1,81%	8,20%
forbes																3,23%	17,09%	1,49%	5,77%
tumba																	15,65%	3,94%	5,92%
dknews																		0,93%	3,76%
ia-centr																			9,60%

Рисунок 2.16 – Анализ текстов статей с различных информационных порталов на сходство при 2 словах в шингле

	zakon	bnews	informburo	today	kstnews	tengrinews	lada.kz	elorda	mix.tn	ru.sputn	kp	alau	baribar	caravan	forbes	tumba	dknews	ia-centr	mgorod
zakon		1,06%	0,62%	0,00%	61,25%	2,24%	19,38%	1,55%	3,23%	0,00%	0,00%	0,00%	58,02%	14,81%	0%	46,60%	14,55%	0,00%	0,00%
bnews			0,97%	0,43%	1,16%	2,24%	0,50%	1,14%	2,96%	1,10%	0,56%	0,68%	0,58%	0,43%	1,52%	2,07%	0,58%	0,59%	0,50%
informburo				1,48%	0,68%	2,07%	0,58%	0,67%	2,80%	1,30%	0,67%	8,04%	0,68%	1,99%	0,58%	1,19%	1,40%	7,46%	9,32%
today					0,00%	7,54%	0,51%	17,57%	8,92%	1,12%	1,74%	3,57%	0,59%	0,44%	8,84%	1,04%	1,81%	0,60%	1,52%
kstnews						2,40%	19,83%	1,75%	3,67%	0,00%	0,00%	0,00%	61,43%	16,33%	0,00%	51,11%	14,43%	0,00%	0,00%
tengrinews							1,71%	11,40%	34,18%	0,45%	0,47%	0,54%	1,92%	2,27%	2,59%	3,07%	2,44%	0,48%	1,26%
lada.kz								1,44%	2,99%	0,00%	0,71%	0,00%	22,12%	13,29%	0,00%	26,77%	13,33%	0,00%	0,00%
elorda									25,00%	0,00%	0,00%	0,00%	1,77%	2,37%	4,44%	2,22%	1,80%	0,00%	1,41%
mix.tn										0,83%	0,87%	1,18%	3,70%	3,66%	4,55%	4,65%	3,77%	0,93%	2,17%
ru.sputniknews											7,76%	7,95%	0,00%	0,00%	0,68%	1,41%	0,00%	0,85%	5,63%
kp												8,43%	0,00%	0,00%	1,43%	1,46%	0,00%	0,89%	5,07%
alau													1,18%	0,00%	0,90%	0,93%	0,00%	12,16%	13,86%
baribar														16,44%	0,00%	45,16%	17,02%	0,00%	0,71%
caravan															0,00%	12,87%	9,09%	0,00%	3,65%
forbes																0,00%	13,33%	0,00%	1,21%
tumba																	10,83%	0,76%	0,61%
dknews																		0,00%	0,72%
ia-centr																			8,66%

Рисунок 2.17 – Анализ текстов статей с различных информационных порталов на сходство при 3 словах в шингле

	zakon	bnews	informburo	today	kstnews	tengrine	lada.kz	elorda	mix.tn	ru.sputn	kp	alau	baribar	caravan	forbes	tumba	dknews	ia-centr	mgorod
zakon		0,00%	0,00%	0,00%	55,95%	1,33%	18,05%	0,76%	2,36%	0,00%	0,00%	0,00%	54,76%	13,33%	0%	44,34%	12,39%	0,00%	0,00%
bnews			0,00%	0,00%	0,00%	0,74%	0,00%	0,57%	1,16%	0,00%	0,00%	0,00%	0,00%	0,00%	0,50%	0,00%	0,00%	0,00%	0,00%
informburo				0,00%	0%	0,00%	0,00%	0,00%	1,38%	0,00%	0,00%	7,14%	0,00%	0,99%	0,00%	0,00%	0,00%	6,72%	6,02%
today					0,00%	4,63%	0,00%	14,47%	5,52%	0,56%	0,57%	0,69%	0,00%	0,00%	5,91%	0,00%	1,20%	0,00%	0,00%
kstnews						1,44%	17,65%	0,88%	2,73%	0,00%	0,00%	0,00%	54,79%	14,86%	0,00%	47,83%	11,11%	0,00%	0,00%
tengrinews							1,28%	8,67%	33,54%	0,00%	0,00%	0,00%	1,44%	1,52%	1,29%	1,30%	1,46%	0,00%	0,00%
lada.kz								0,71%	2,21%	0,00%	0,00%	0,00%	20,87%	12,00%	0,00%	24,62%	11,48%	0,00%	0,00%
elorda									17,53%	0,00%	0,00%	0,00%	0,88%	1,18%	2,96%	0,74%	0,91%	0,00%	0,00%
mix.tn										0,00%	0,00%	0,00%	2,75%	2,42%	2,24%	2,27%	2,83%	0,00%	0,00%
ru.sputniknews											5,08%	4,49%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	4,20%
kp												4,71%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	3,57%
alau													0,00%	0,00%	0,00%	0,00%	0,00%	10,96%	10,68%
baribar														14,97%	0,00%	42,11%	14,74%	0,00%	0,00%
caravan															0,00%	10,98%	7,10%	0,00%	1,54%
forbes																0,00%	12,61%	0,00%	0,00%
tumba																	9,09%	0,00%	0,00%
dknews																		0,00%	0,00%
ia-centr																			7,87%

Рисунок 2.18 – Анализ текстов статей с различных информационных порталов на сходство при 4 словах в шингле

При изучении полученных результатов можно заметить, как значительно понижается процент пересечений шинглов между текстами с повышением числа слов в одном шингле.

Как раз на рисунках 2.15 и 2.18 можно заметить, как уменьшилось количество пересечений и это при учете того что с повышением количества слов в шингле повышалась и значимость пересечений на каждом новом уровне на 2%. В результате мы можем увидеть матрицу, в которой зеленым цветом отмечены результаты сходства, которые считаются значительными, а желтым незначительные. При это все значения, которые отмечены серым цветом исключаются из анализа, так как процент сходства между текстами недостаточен для заключения вывода о том, что имеется заимствование.

Матрицы в данном случае будут симметричны относительно оси, которая показывает полное сходство между одинаковыми наборами текстов. Например текст взятый с tengrinews будет полностью идентичен (процент пересекающихся шинглов 100%) точно такому же тексту с tengrinews.

При проведении анализа с помощью алгоритма шинглов необходимо учитывать, что количество пересечений между текстом А и Б одинаковое, как и в случае сравнения текстов Б и А. Следовательно, полученная матрица будет симметрична, и достаточно рассмотреть только ее часть.

Теперь, когда мы получили результаты касательно времени появления статей в сети Интернет, а также составили матрицу сходства между ними, возможно приступать к последнему этапу анализа: построению картины диффузии информации и обработке результатов.

3 Результаты анализа данных и картина диффузии информации среди средств массовой информации в РК

В ходе предыдущих этапов исследования были последовательно изучены несколько факторов влияющих на процесс распространения информации.

Первый из них это количество агентов распространения информации, которые проживают в том или ином регионе. Благодаря данным о количестве проживающего населения, его демографическом составе, уровне компьютерной грамотности и интернет покрытия удалось определить, как общее количество потенциальных участников распространения информации, так и построить распределение согласно регионам.

После того как удалось выяснить общий размер сети, стало необходимо определить на основе каких параметров можно оценить, что информационный ресурс воздействует на эту аудиторию. Для этого был проведен анализ рынка средств массовой информации, изучены такие параметры, как количество просмотров, количество посетителей, пользовательские сессии, количество новостей, число переходов на каждую новость. И теперь зная общий размер сети, а также абсолютные показатели каждого из СМИ стало возможным определить рейтинг 25 самых востребованных информационных ресурсов.

Как следующий этап было проведено исследование 25 лучших СМИ за 2019 год, были проанализированы показатели, которые влияют на процесс диффузии информации. И наконец был сформирован список ресурсов, которые были популярны у населения в указанный период, и при этом находились в этом рейтинге достаточно долго, чтобы не считать, что это было влияние только одной информационной компании. В дальнейшем эти ресурсы были соотнесены с их географическим местоположением и количеством населения, которые проживают в этих регионах.

Далее был проведен сбор датасета для анализа, а именно упоминания на указанных ресурсах о прекращении полномочий первого президента РК. Для этого были задействованы инструменты и сервисы по выявлению контента по строго заданным источникам, а также алгоритмы поиска данных о дате первой публикации статей в сети Интернет.

После того как датасет был сформирован последним этапом в обработке данных стало выявление процента сходства между статьями. Это было сделано для того чтобы понять насколько велика вероятность что один из ресурсов заимствовал информацию у другого, а, следовательно, происходил процесс диффузии информации. Результатом стали 4 матрицы, построенные с помощью алгоритма выявления дубликатов и плагиата – алгоритма шинглов. Причем статистически значимой из них являлась только матрица, построенная с учетом что в одном шингле содержится 4 слова.

Все это позволило сформировать картину распространения информации 19 марта 2019 года, во время телевизионного обращения первого президента Республики Казахстан – Назарбаева Н.А. между ключевыми СМИ на рисунке 3.1.

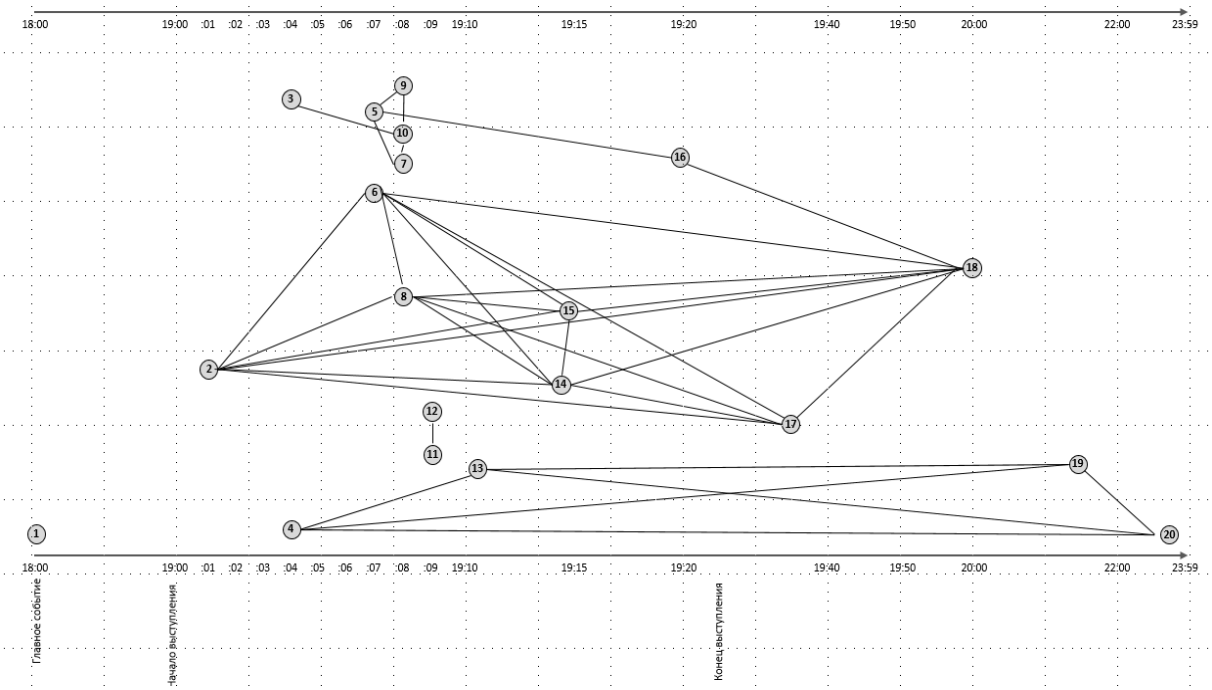

Рисунок 3.1 – Процесс диффузии информации между СМИ во время телевизионного обращения главы Республики Казахстан

В итоге получился граф, где узлами выступают СМИ, порядковые номера которым присвоены в зависимости от времени публикации статьи на изучаемую тематику. Соответствие СМИ можно увидеть в таблице 3.1:

Таблица 3.1 – Распределение СМИ в зависимости от времени публикации статьи

Номер в графе	**Ресурс**	**Дата публикации**
2	zakon	19.03.2019,19:01
3	bnews	19.03.2019,19:04
4	informburo	19.03.2019,19:04
5	today	19.03.2019,19:07
6	kstnews	19.03.2019,19:07
7	tengrinews	19.03.2019,19:08
8	lada.kz	19.03.2019,19:08
9	elorda	19.03.2019,19:08
10	mix.tn	19.03.2019,19:08
11	ru.sputniknews	19.03.2019,19:09
12	kp	19.03.2019,19:09
13	alau	19.03.2019,19:11
14	baribar	19.03.2019,19:15
15	caravan	19.03.2019,19:15
16	forbes	19.03.2019,19:20
17	tumba	19.03.2019,19:37
18	dknews	19.03.2019,20:00
19	ia-centr	19.03.2019,21:32

Как видно из получившегося графа, он состоит из нескольких кластеров, некоторые из которых представляют из себя полный граф из-за наличия большого сходства в тексте между каждой парой ресурсов.

Однако в данном случае еще имеет влияние и время появления статьи на ресурсе, и, следовательно, чем раньше появилась информация на ресурсе, тем выше его значимость в общей сети. На рисунке 3.2 показан кластер, распространение информации в котором началось с самого первого упоминания в СМИ новости о прекращении полномочий президента, а именно с ресурса zakon.kz. Как результат в данном кластере, который представляет из себя 6 связанных между собой ресурсов именно zakon.kz будет иметь наивысшую значимость, так как новость впервые появилась именно там, а процент сходства с ресурсами разместившими информацию позже все достаточно высок.

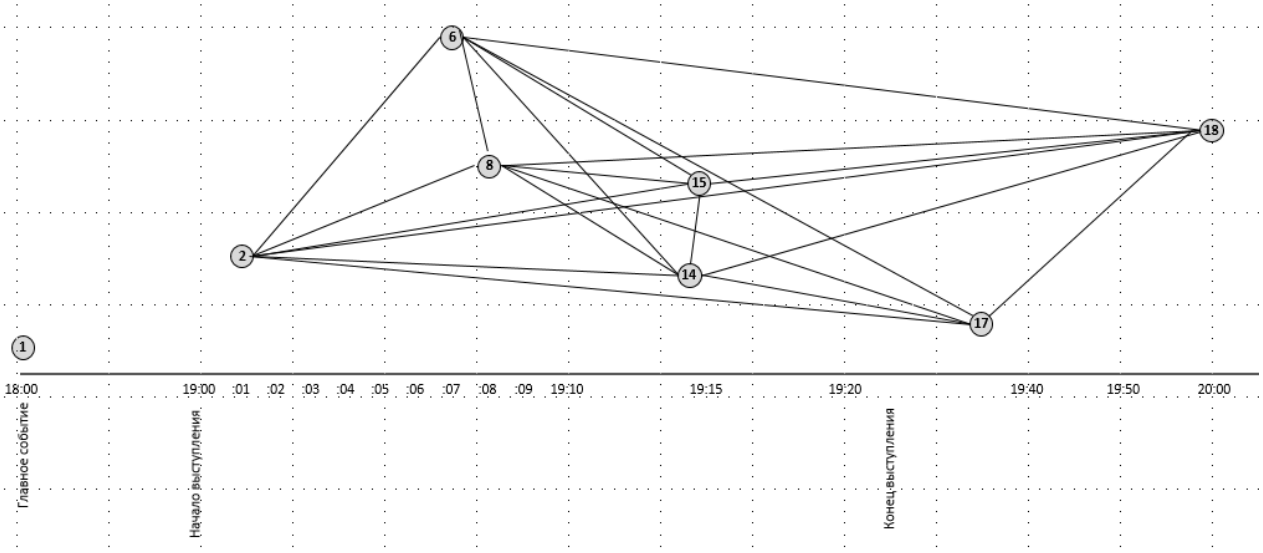

Рисунок 3.2 – Кластер СМИ распространение информации в котором началось с ресурса zakon.kz

Следующий кластер представляет из себя следующую ситуацию: ресурс bnews.kz разместил новость раньше остальных, при этом он обладает с высоким процентом схожести между группой и 4 ресурсов, разместивших информацию с разницей в 1 минуту. При этом у этом эти ресурсы представляют из себя полный граф. Позже наблюдалось распространение информации и на другие ресурсы, разместившими информацию сравнительно поздно. При этом картина распределения информации представляет из себя аналогичную картине принятия инноваций. Сначала имеется новатор, который находится меньшинстве, в дальнейшем начинают появляться ранние последователи (наблюдается резкий рост количества участников) а затем идет фаза принятия инновации, и появления поздних последователь. Более подробно этот процесс рассматривался в главе 1.2.2 на рисунке 1.3.

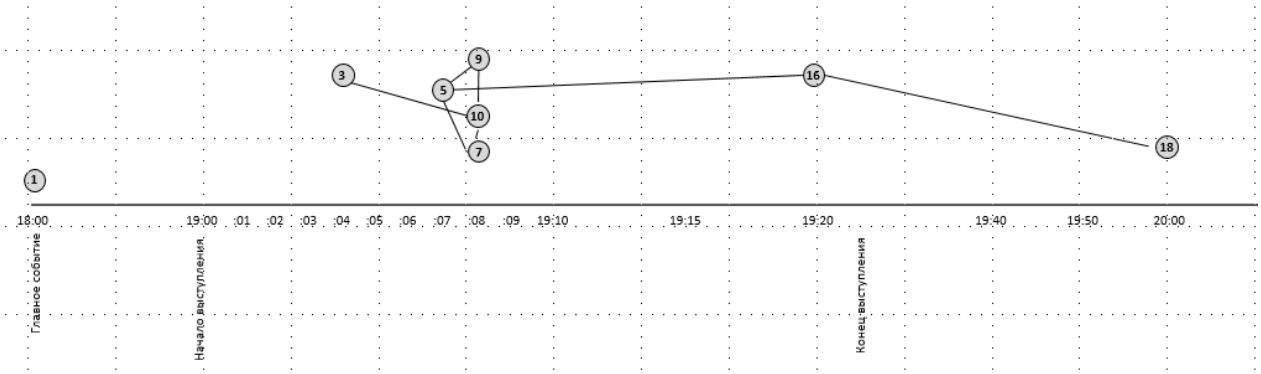

Рисунок 3.3 – Кластер СМИ распространение информации в котором началось с ресурса bnews.kz

Последний крупный кластер из 4 средств массовой информации представляет еще один полный граф, причем распространение информации в нем велось практически линейно от одного ресурса к другому, если принимать во внимание уровень заимствования между ресурсами, от 4 к 13, от 13 к 19 и к 20. При этом учитывая, что нельзя точно зафиксировать точное авторство какому-либо из источников, остается только основываться на времени появления статьи на ресурсе, а в нашем случае это informburo.kz

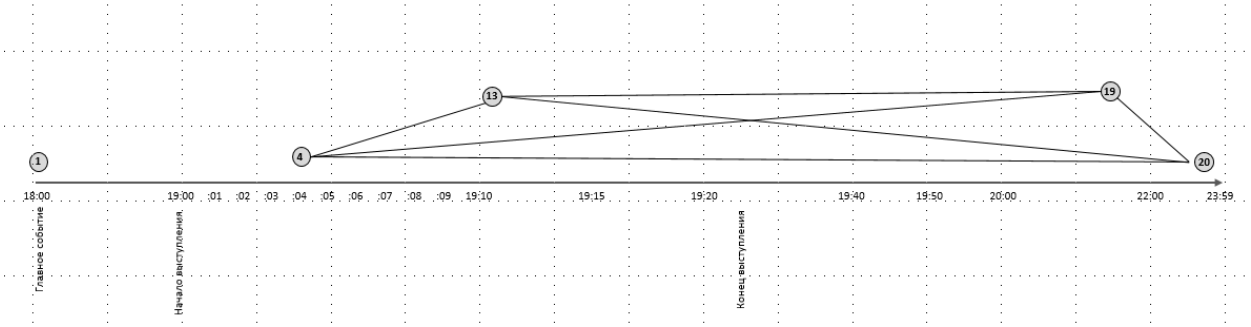

Рисунок 3.4 – Кластер СМИ распространение информации в котором началось с ресурса informburo.kz

Последний кластер представляет пару ресурсов, разместивших информацию практически одновременно. Здесь невозможно полностью воссоздать картину распространения информации, т.к. есть две гипотезы: или ресурсы имеют одного автора, который размещает информацию на обоих одновременно, или они имеют некоторый третий первоисточник, из которого заимствована информация.

Соответственно ресурсы ru.sputniknews и kp.kz имеют установленную взаимосвязь, однако доподлинно определить первоисточник не удалось. Есть еще одна гипотеза, что на эту пару могло повлиять то что изначальная редакция одного и другого СМИ находится в Российской Федерации и первичный источник информации находится там, а так как мы проводили анализ СМИ ведущих свою деятельность только на территории РК, то упустили этот источник из виду.

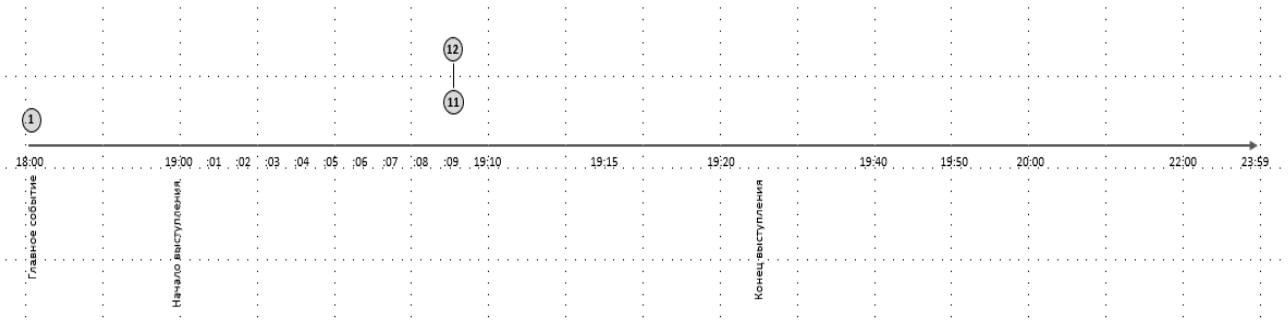

Рисунок 3.5 – Кластер СМИ распространение информации в котором началось с ресурса informburo.kz

В результате можно сказать что с помощью методов и подходов анализа информации и ее распространения, которые были описаны в 1 главе удалось определить взаимосвязи между источниками, а благодаря анализу даты появления публикации в сети Интернет, удалось построить распределение этих источников по временной шкале.

Все это позволило воссоздать картину распространения информации о прекращении полномочий первого президента Республики Казахстан Назарбаева Н.А. среди отечественных СМИ. Как уже было упомянуто ранее, данный инфоповод был выбран, ввиду того что достоверно можно было установить какое ресурс является главным первоисточником, а именно: akorda.kz, на котором было размещено обращение к народу.

В ходе исследований удалось выяснить что сама страница с обращением была создана ровно за час до начала обращения. А первое упоминание этой новости в СМИ (zakon.kz) произошло всего через минуту после его публикации, что говорит об автоматическом характере анализа новостей, с государственного ресурса.

Так как методология исследования в ходе выполнения данной диссертационной работы была отработана на 25 СМИ, следующими шагами в исследовании могут стать: автоматизация процесса сбора информации о статьях с более чем 25 источников, автоматическое формирование сети распространения информации на определенную тематику и в дальнейшем предоставление данной услуги как сервиса.

При этом на вход при полной автоматизации может подаваться запрос на тематику или бренд. Далее будет происходить процесс поиска и сбор датасета с различных источников. После чего он, пройдя обработку с помощью алгоритма на основе шинглов позволит сформировать граф диффузии информации в распределенной системе. Все это в дальнейшем позволит создать более точный инструмент для аналитики успешности распространения информации, по сравнению с существующими решениями, которые останавливаются на этапе поиска и сбора набора данных об упоминаниях в распределенной сети.

Заключение

Как было замечено в 1 главе данной работы, распространение информации относится к тому, как мнения (состояния) отдельных узлов в сети (графе) эволюционируют со временем. Два явления, которые вызывают распространение информации в распределенных сетях: первое это – передача и принятие информации, а второе - гомофилия. Диффузия на основе «заражения» определяются влиянием соседей, тогда как диффузия на основе гомофилии определяется свойствами узлов (которые коррелируют между соседями). Динамические модели для таких процессов распространения информации в распределенных сетях таких как распространение новостей, инноваций и так далее, позволяют воссоздать реальную картину диффузии информации.

Ранее были упомянуты несколько исследований таких крупных сетей как Facebook, Twitter и процессы распространения информации в других социальных сетях. Однако исследования диффузии среди традиционных СМИ достаточно редки. Именно в этом и заключается научная новизна проведенного исследования процессов распространения информации в таких распределенных системах как республиканские СМИ.

В итоге удалось выяснить как различаются между собой типы связей между агентами распространения информации, какие могут быть типы передаваемых сообщений, как выглядит динамика распространения, а также были описаны потенциальные пути распространения информации. Все это позволило приступить к поиску наиболее эффективных методов анализа текстовой информации, благодаря исследованиям в области анализа фейковых новостей. Методы борьбы с распространением недостоверной информации, также дали ценные данные для построения модели диффузии в распределенной системе. И последним теоретическим исследованием стал анализ существующих решений в этой области.

Во время практического исследования удалось достигнуть следующих результатов: была проведена оценка размера общей сети распространения информации в зависимости от региона РК, проведена аналитика всего рынка СМИ для определения доли каждого из ресурсов в общем информационном пространстве, выявлены потенциальные наиболее влиятельные СМИ, а также параметры их оценки. Указанные СМИ были соотнесены с регионом их работы, а также был сформирован датасет с упоминаниями по определенной тематике. Все это позволило в конечном итоге определить взаимосвязи между источниками и сформировать общую картину распространения информации в распределенной системе. Анализ самого графа диффузии информации в качестве результата позволил определить главных участников во всей сети.

Дальнейшими шагами этого исследования могут выступить разработки по автоматизации процесса анализа: подключение к источникам данных для быстрого формирования датасета, автоматическое выявление взаимосвязей и как результат, формирование графа диффузии информации по запросу пользователя.

Список использованных источников

[1] Granovetter, M. The strength of weak ties. – The American J. of Sociology 78.

[2] Энциклопедическое определение термина информация // свободный доступ по ссылке https://www.dictionary.com/browse/information , (дата запроса 10.05.2020)

[3] Luciano Floridi. Information - A Very Short Introduction. / Oxford University Press 2010.

[4] Kaye, J., Levitt, M., Nevins, J., Golden, J., Schmitt V. Communication intimacy one bit at time. - CHI Extended Abstracts on Human Factors in Computing Systems, 2005.

[5] Xiao, Z., Guo, L., Tracey, J. Understanding instant messaging traffic characteristics. - Proc. ICDCS, 2007

[6] Golder S., Wilkinson D., Huberman B. Rhythms of social interaction: Messaging within a massive online network. - 3rd International Conference on Communities and Technologies, 2007.

[7] Boguna, M., Krioukov, D., and Claffy, K. C. Navigability of complex networks., сентябрь 2007.

[8] Kossinets, G., Kleinberg, J., Watts, D. The structure of information pathways in a social communication network. – KDD, Las-Vegas 2008

[9] Gibson, D. Concurrency and commitment: Network scheduling and its consequences for diffusion. J. of Mathematical Sociology 29, 2005.

[10] Gruhl, D., Liben-Nowell, D., Guha, R., and Tomkins, A. Information diffusion through blogosphere. 13th International World Wide Web Conference, 2004.

[11] Lewis D. Convention: a Philosophical Study. Harvard University Press, Cambridge, MA, 1969

[12] Harary F., Norman R., Cartwright D. Structural Models. Wiley, New York, 1965.

[13] Chwe, M. S.-Y. Communication and coordination in social networks. Rev. of Economic Studies 67, 2000.

[14] Kermack W., and McKendrick, A. A contribution to the mathematical theory of epidemics. Roy.Soc. Lond. A 115, 1927.

[15] Kitsak, M., Gallos, L., Havlin, S., Liljeros, F.,Muchnik, L., Stanley, H., and Makse, H. Identifying influential spreaders in complex networks., январь 2010

[16] Freeman, L. Centrality in social networks: Conceptual clarification. Social Networks 1, 1979.

[17] Emanuelson, P., and Willer, D. One-shot ex-change networks and the shadow of the future. Social Networks 31, 2009.

[18] Leskovec, J., and Horvitz, E. Planetary-scale views on a large instant-messaging network. 17th international conference on World Wide Web, New York, NY, USA, 2008.

[19] Milgram, S. The small world problem. Psychology Today 1, 1967

[20] Wilson, C., Boe, B., Sala, A., Puttaswamy, K.,and Zhao, B. User interactions in social networks and their implications. Proceedings of the 4th ACM European conference on Computer systems New York, NY, USA, 2009.

[21] Lindelauf, R., Borm, P., and Hamers, H. The influence of secrecy on the communication structure of covert networks. Social Networks 31, 2009

[22] Richard Smith. What is Digital Media. Свободный доступ по ссылке: https://thecdm.ca/news/what-is-digital-media (дата запроса 08.05.2020)

[23] Southwell, B. G., Yzer, M. C. The roles of interpersonal communication in mass media campaigns. Communication Yearbook, 31, 2007.

[24] Rössler, P. The individual agenda-designing process: How interpersonal communication, egocentric networks, and mass media shape the perception of political issues by individuals. Communication Research 26, 1999

[25] Schmitt-Beck, R. Interpersonal communication. C. Holtz-Bacha & L. L. Kaid (Eds.), Encyclopedia of political communication. Los Angeles, CA: Sage, 1999

[26] Gehrau, V., Döveling, K., Sommer, D., Dunlop, S. Antagonistic and synergetic impacts of conversation on nonpersuasive media effects. Communication Research, 41(4), 2014.

[27] Lull, J.. Inside family viewing: Ethnographic research on television's audiences. New York, NY: Routledge, 1999

[28] Sommer, D.. Media effects, interpersonal communication and beyond: An experimental approach to study conversations about the media and their role in media reception. Journal for Communication Studies, 6(1), 2013.

[29] Greenberg, S. R.. Conversations as units of analysis in the study of personal influence. Journalism Quarterly, 52(1), 1975

[30] Kepplinger, H. M.. Media effects: Direct and indirect effects. In W. Donsbach , The international encyclopedia of communication. Oxford, UK: Wiley Blackwell 2008.

[31] Vu, H. N. N., & Gehrau, V. Agenda diffusion: An integrated model of agenda setting and interpersonal communication. Journalism and Mass Communication Quarterly, 87(1), 2010.

[32] Robinson, J. P., & Levy, M. R.. Interpersonal communication and news comprehension. Public Opinion Quarterly, 50(2), 1986.

[33] Southwell, B. G. 2005. Between messages and people: A multilevel model of memory for television content. Communication Research, 32(1), 2005

[34] O'Sullivan, P. B.. Masspersonal communication: Rethinking the mass-interpersonal divide. Paper presented at the annual meeting of the International Communication Association, New York, 2005

[35] Young, S. D., Belin, T. R., Klausner, J., & Valente, T. W. . Methods for measuring diffusion of a social media-based health intervention. Social Networking, 4(2), 2015

[36] Ziegele, M., & Quiring, O.. Conceptualizing online discussion value: A multidimensional framework for analyzing user comments on mass-media websites. Communication Yearbook, 37, 2013.

[37] Berger, J. Beyond viral: Interpersonal communication in the Internet age. Psychological Inquiry, 24(4), 2013

[38] Petric, G., Petrov A.,Vehovar, V. Social uses of interpersonal communication ˇ technologies in a complex media environment. European Journal of Communication, 26(2), 2011.

[39] Podschuweit, N.. Interpersonal Communication: Media Influence on. The International Encyclopedia of Media Effects, 2017

[40] Nicole A Cooke. 2017. Posttruth, truthiness, and alternative facts: Information behavior and critical information consumption for a new age. The Library Quarterly 87, 3 2017.

[41] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Journal of Economic Perspectives 31, 2 2017.

[42] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu.. Fake news detection on social media: A data mining perspective. Explorations Newsletter 19, 1, 2017

[43] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2018. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans, 2018.

[44] Srijan Kumar and Neil Shah. False information on web and social media: 2018

[45] Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B Everett, et al. Fake News vs Satire: A Dataset and Analysis. 10th ACM Conference on Web Science. 2018

[46] Claire Wardle. Fake news. First Draft News 2017.

[47] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. ACM Computing Surveys (CSUR) 51, 2 2018.

[48] Udo Undeutsch.. Courtroom evaluation of eyewitness testimony. Applied Psychology 33, 1 1984.

[49] Glynis Bogaard, Ewout H Meijer, Aldert Vrij, and Harald Merckelbach. Scientific Content Analysis (SCAN) cannot distinguish between truthful and fabricated accounts of a negative Event. Frontiers in psychology 7 2016.

[50] Galit Nahari, Aldert Vrij, and Ronald P Fisher. Does the truth come out in the writing? Scan as a lie detection tool. Law and Human Behavior 36, 1 2012.

[51] Lina Zhou, Judee K Burgoon, Jay F Nunamaker, and Doug Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. Group decision and negotiation 13, 1 2004.

[52] David B Buller and Judee K Burgoon. Interpersonal deception theory. Communication theory 6, 3 1996.

[53] David B Buller, Judee K Burgoon, Aileen Buslig, and James Roiger. 1996. Testing interpersonal deception theory: The language of interpersonal deception. Communication theory 6, 3 1996.

[54] James W Pennebaker. Linguistic inquiry and word count: LIWC 2001.

[55] Gary D Bond and Adrienne Y Lee. Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. Applied Cognitive Psychology 19, 3 2005.

[56] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. Personality and social psychology bulletin 29, 5 2003.

[57] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In Proceedings of the 20th international conference on World wide web. 2011

[58] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In International Conference on Social Informatics. Springer, 2014.

[59] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, 2012

[60] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2016

[61] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: Recognizing clickbait as false news. In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection. 2015

[62] Myle Ott, Claire Cardie, and Jeffrey T Hancock. Negative Deceptive Opinion Spam.. In HLT-NAACL. 2013

[63] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

[64] Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan, and Randolph Quirk. Longman grammar of spoken and written English. Vol. 2. MIT Press Cambridge, MA, 1999

[65] Paul Rayson, Andrew Wilson, and Geoffrey Leech. 2001. Grammatical word class variation within the British National Corpus sampler. Language and Computers 36, 1, 2001.

[66] Mark Johnson. PCFG models of linguistic tree representations. Computational Linguistics 24, 4 1998.

[67] Rada Mihalcea and Carlo Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Association for Computational Linguistics, 2009

[68] Phil Blunsom, Edward Grefenstette, and Nal Kalchbrenner. A convolutional neural network for modelling sentences. In Proceedings of the 52nd

Annual Meeting of the Association for Computational Linguistics. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014.

[69] Yoon Kim. Convolutional neural networks for sentence classification, 2014

[70] Wen-tau Yih, Xiaodong He, and Christopher Meek. Semantic parsing for single-relation question answering. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2014

[71] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition, 1998

[72] William Yang Wang.. “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, 2017

[73] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space, 2013

[74] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams. Learning representations by back-propagating errors. Cognitive modeling 5, 3 1988

[75] Sahil Chopra, Saachi Jain, and John Merriman Sholar. Towards Automatic Identification of Fake News: HeadlineArticle Stance Detection with LSTM Attention Models, 2017

[76] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, Yan Liu, University of Southern California, Combating Fake News: A Survey on Identification and Mitigation Techniques, 2019

[77] Сайт компании IMAS, свободный доступ по ссылке https://imas.kz/#preim (дата запроса 01.05.2020)

[78]. Губанов Д.А., Информационные процессы в социальных сетях (на примере сети Хабрахабр) // Интернет-конференция по проблемам управления.: ИПУ РАН. – 2010. – С.1 – 2.

[79]. Барабанов И.Н., Коргин Н.А., Новиков Д.А., Чхартишвили А. Г. Динамические модели информационного управления в социальных сетях // Автомат. и телемех., – 2010. – № 11 – С. 150–162.

[80]. Министерство национальной экономики Республики Казахстан Комитет по статистике // Данные по составу и численности населения 2018 год. Свободный доступ по ссылке https://stat.gov.kz/official/dynamic (дата обращения: 10.02.2020)

[81]. Министерство национальной экономики Республики Казахстан Комитет по статистике // Данные по уровню компьютерной грамотности населения за 2018 год. Свободный доступ по ссылке https://stat.gov.kz/official/dynamic (дата обращения: 10.02.2020)

[82]. Информационный портал Tengrinews // Когда 5G появится в регионах Казахстана – дата публикации 01.10.2019 – Свободный доступ по ссылке: https://tengrinews.kz/kazakhstan_news/kogda-5g-poyavitsya-v-regionah-kazahstana-380439/ (дата обращения: 10.02.2020)

[83]. Министерство информации и общественного развития Республики Казахстан // Статистические данные по составу населения, использующему интернет. Свободный доступ по ссылке: http://qogam.gov.kz/ru (дата обращения: 10.02.2020)

[84] Портал Activator consulting, All You Need to Know About Media Efficiency, свободный доступ по ссылке: https://activatorconsulting.com/all-you-need-to-know-about-media-efficiency/ (дата запроса 03.05.2020)

[85] Mandelli, A., Accoto, C., & Mari, A.. Social media metrics: Practices of measuring brand equity and reputation in online social collectives. Proceedings of the 6th International Conference Thought Leaders in Brand Management, Lugano, Switzerland, 2010

[86] Philip M. Napoli Professor School of Communication Information Rutgers University Winter measuring media impact an overview of the field, 2013.

[87] Статистика по рынку СМИ РК, свободный доступ по ссылке: https://www.liveinternet.ru/stat/kz/media/ (дата запроса 02.03.2020)

[88] Статистика по 25 топовым СМИ, свободный доступ по ссылке https://mediametrics.ru/top_smi/?kz:all:2019-12:domains (дата запроса 02.03.2020)

89 Сервис для определения даты размещения веб страницы в сети Интернет, свободный доступ по ссылке http://carbondate.cs.odu.edu/ (дата запроса 02.03.2020)

90 Описание работы алгоритма шинглов для анализа схожести нескольких текстов. Свободный доступ по ссылке https://habr.com/ru/company/antiplagiat/blog/445952/ (дата запроса 02.03.2020)

Приложение А

Статистические данные по 25 самым популярным СМИ в РК

Место	Издание	Переходы	Новость	Переходы/ Новость	Доля	Дата
1	tengrinews	679,286	2297	296	36.45%	янв.19
2	alau	183,606	565	325	9.85%	янв.19
3	kstnews	139,045	391	356	7.46%	янв.19
4	lada	120,169	471	255	6.45%	янв.19
5	zakon	104,348	3300	32	5.60%	янв.19
6	informburo	100,262	1380	73	5.38%	янв.19
7	qostanay	96,301	748	129	5.17%	янв.19
8	bnews	83,581	1094	76	4.48%	янв.19
9	uralskweek	69,349	202	343	3.72%	янв.19
10	dknews	56,522	2836	20	3.03%	янв.19
11	mgorod	49,979	660	76	2.68%	янв.19
12	ng	30,479	272	112	1.64%	янв.19
13	kaz.tengrinews	27,751	632	44	1.49%	янв.19
14	tumba	25,357	351	72	1.36%	янв.19
15	altyn-orda	22,496	278	81	1.21%	янв.19
16	ia-centr	14,701	134	110	0.79%	янв.19
17	elorda	11,451	273	42	0.61%	янв.19
18	ru.sputniknews	10,89	1228	9	0.58%	янв.19
19	forbes	8717	628	14	0.47%	янв.19
20	mix.tn	6181	90	69	0.33%	янв.19
21	baribar	5223	159	33	0.28%	янв.19
22	vecher	3446	1	3446	0.18%	янв.19
23	kapshagai	1795	51	35	0.10%	янв.19
24	kapital	1703	195	9	0.09%	янв.19
25	aktobegazeti	1558	169	9	0.08%	янв.19
1	tengrinews	698,977	2328	300	38.05%	фев.19
2	alau	151,049	563	268	8.22%	фев.19
3	lada	132,284	491	269	7.20%	фев.19
4	informburo	107,172	1294	83	5.83%	фев.19
5	zakon	105,246	3166	33	5.73%	фев.19
6	kstnews	100,507	344	292	5.47%	фев.19
7	qostanay	84,525	744	114	4.60%	фев.19
8	uralskweek	78,266	191	410	4.26%	фев.19
9	bnews	69,079	1329	52	3.76%	фев.19
10	mgorod	62,447	575	109	3.40%	фев.19
11	altyn-orda	43,006	454	95	2.34%	фев.19
12	kaz.tengrinews	41,571	726	57	2.26%	фев.19
13	dknews	26,798	2305	12	1.46%	фев.19

14	elorda	26,725	392	68	1.45%	фев.19
15	ng	25,52	299	85	1.39%	фев.19
16	tumba	24,957	393	64	1.36%	фев.19
17	ia-centr	15,021	119	126	0.82%	фев.19
18	forbes	9507	596	16	0.52%	фев.19
19	ru.sputniknews	9152	1189	8	0.50%	фев.19
20	baribar	6153	223	28	0.33%	фев.19
21	mix.tn	4284	92	47	0.23%	фев.19
22	timeskz	1769	104	17	0.10%	фев.19
23	kp	1692	182	9	0.09%	фев.19
24	kapshagai	1576	25	63	0.09%	фев.19
25	dixinews	1388	262	5	0.08%	фев.19
1	tengrinews	638,13	2351	271	38.95%	мар.19
2	lada	130,654	515	254	7.97%	мар.19
3	alau	116,855	469	249	7.13%	мар.19
4	zakon	107,84	3160	34	6.58%	мар.19
5	informburo	100,503	1384	73	6.13%	мар.19
6	qostanay	73,551	727	101	4.49%	мар.19
7	mgorod	67,015	568	118	4.09%	мар.19
8	kstnews	62,409	513	122	3.81%	мар.19
9	altyn-orda	44,32	434	102	2.71%	мар.19
10	bnews	42,89	1202	36	2.62%	мар.19
11	dknews	40,695	2745	15	2.48%	мар.19
12	uralskweek	37,72	135	279	2.30%	мар.19
13	elorda	29,72	372	80	1.81%	мар.19
14	ng	29,245	304	96	1.78%	мар.19
15	kaz.tengrinews	27,357	758	36	1.67%	мар.19
16	ia-centr	24,102	141	171	1.47%	мар.19
17	tumba	22,242	365	61	1.36%	мар.19
18	ru.sputniknews	12,277	1286	10	0.75%	мар.19
19	forbes	9207	623	15	0.56%	мар.19
20	mix.tn	4561	74	62	0.28%	мар.19
21	baribar	1961	198	10	0.12%	мар.19
22	smirnovo	1868	61	31	0.11%	мар.19
23	kp	1760	172	10	0.11%	мар.19
24	dixinews	1670	303	6	0.10%	мар.19
25	voxpopuli	1463	86	17	0.09%	мар.19
1	tengrinews	741,486	2562	289	40.10%	апр.19
2	zakon	185,82	3363	55	10.05%	апр.19
3	lada	128,199	551	233	6.93%	апр.19
4	informburo	95,163	1447	66	5.15%	апр.19
5	alau	90,095	471	191	4.87%	апр.19
6	bnews	89,869	1351	67	4.86%	апр.19

7	qostanay	85,295	979	87	4.61%	апр.19
8	kstnews	78,899	448	176	4.27%	апр.19
9	mgorod	71,181	594	120	3.85%	апр.19
10	altyn-orda	45,544	569	80	2.46%	апр.19
11	dknews	40,178	2537	16	2.17%	апр.19
12	uralskweek	35,655	149	239	1.93%	апр.19
13	ng	31,049	344	90	1.68%	апр.19
14	elorda	24,071	488	49	1.30%	апр.19
15	tumba	22,232	377	59	1.20%	апр.19
16	kaz.tengrinews	19,569	718	27	1.06%	апр.19
17	forbes	18,579	700	27	1.00%	апр.19
18	ru.sputniknews	10,945	1341	8	0.59%	апр.19
19	ia-centr	9440	144	66	0.51%	апр.19
20	mix.tn	3947	91	43	0.21%	апр.19
21	baribar	2731	249	11	0.15%	апр.19
22	aktobegazeti	2711	180	15	0.15%	апр.19
23	voxpopuli	2551	74	34	0.14%	апр.19
24	online.zakon	2331	394	6	0.13%	апр.19
25	kp	1734	218	8	0.09%	апр.19
1	tengrinews	873,548	2518	347	44.77%	май.19
2	zakon	190,369	3013	63	9.76%	май.19
3	lada	148,417	564	263	7.61%	май.19
4	bnews	127,331	1062	120	6.53%	май.19
5	alau	101,542	524	194	5.20%	май.19
6	informburo	96,818	1424	68	4.96%	май.19
7	mgorod	66,123	586	113	3.39%	май.19
8	qostanay	65,52	812	81	3.36%	май.19
9	kstnews	65,295	365	179	3.35%	май.19
10	uralskweek	36,872	190	194	1.89%	май.19
11	elorda	31,865	453	70	1.63%	май.19
12	ng	29,762	344	87	1.53%	май.19
13	dknews	28,778	2634	11	1.47%	май.19
14	tumba	18,528	324	57	0.95%	май.19
15	forbes	15,944	563	28	0.82%	май.19
16	altyn-orda	10,963	305	36	0.56%	май.19
17	kaz.tengrinews	9791	657	15	0.50%	май.19
18	ru.sputniknews	8127	1237	7	0.42%	май.19
19	ia-centr	6281	134	47	0.32%	май.19
20	toppress	3019	165	18	0.15%	май.19
21	voxpopuli	2557	87	29	0.13%	май.19
22	timeskz	2293	164	14	0.12%	май.19
23	mix.tn	2171	74	29	0.11%	май.19
24	online.zakon	1335	266	5	0.07%	май.19

25	baribar	1124	175	6	0.06%	май.19
1	tengrinews	1,181,394	2647	446	51.16%	июн.19
2	zakon	170,618	2921	58	7.39%	июн.19
3	lada	150,273	568	265	6.51%	июн.19
4	bnews	148,838	1044	143	6.45%	июн.19
5	informburo	126,663	1365	93	5.49%	июн.19
6	alau	107,121	488	220	4.64%	июн.19
7	kstnews	75,727	367	206	3.28%	июн.19
8	mgorod	72,691	532	137	3.15%	июн.19
9	uralskweek	38,82	222	175	1.68%	июн.19
10	toppress	29,722	331	90	1.29%	июн.19
11	ng	29,269	353	83	1.27%	июн.19
12	elorda	26,915	376	72	1.17%	июн.19
13	today	26,844	748	36	1.16%	июн.19
14	dknews	23,135	1997	12	1.00%	июн.19
15	kaz.tengrinews	20,082	789	25	0.87%	июн.19
16	tumba	19,582	351	56	0.85%	июн.19
17	forbes	19,316	551	35	0.84%	июн.19
18	ia-centr	13,01	147	89	0.56%	июн.19
19	ru.sputniknews	11,641	1434	8	0.50%	июн.19
20	shakhty	1923	1	1923	0.08%	июн.19
21	altyn-orda	1898	280	7	0.08%	июн.19
22	mix.tn	1617	63	26	0.07%	июн.19
23	baribar	1599	127	13	0.07%	июн.19
24	online.zakon	1579	255	6	0.07%	июн.19
25	smirnovo	1301	46	28	0.06%	июн.19
1	tengrinews	1,390,371	3145	442	53.54%	июл.19
2	lada	178,537	512	349	6.88%	июл.19
3	informburo	133,729	1390	96	5.15%	июл.19
4	alau	116,812	492	237	4.50%	июл.19
5	bnews	116,489	846	138	4.49%	июл.19
6	zakon	109,611	3206	34	4.22%	июл.19
7	qostanay	93,656	1038	90	3.61%	июл.19
8	kstnews	92,995	372	250	3.58%	июл.19
9	mgorod	69,974	572	122	2.69%	июл.19
10	dknews	40,977	2619	16	1.58%	июл.19
11	kaz.tengrinews	36,021	833	43	1.39%	июл.19
12	today	27,788	1259	22	1.07%	июл.19
13	ng	27,666	290	95	1.07%	июл.19
14	elorda	25,284	308	82	0.97%	июл.19
15	ia-centr	22,272	161	138	0.86%	июл.19
16	uralskweek	21,548	184	117	0.83%	июл.19
17	toppress	20,28	310	65	0.78%	июл.19

18	tumba	19,095	326	59	0.74%	июл.19
19	forbes	14,716	538	27	0.57%	июл.19
20	ru.sputniknews	9217	1248	7	0.35%	июл.19
21	kp	4159	884	5	0.16%	июл.19
22	mix.tn	3847	68	57	0.15%	июл.19
23	inbusiness	3663	223	16	0.14%	июл.19
24	mtrk	2803	159	18	0.11%	июл.19
25	vecher	1921	1	1921	0.07%	июл.19
1	tengrinews	1,211,094	3030	400	48.70%	авг.19
2	bnews	159,524	926	172	6.41%	авг.19
3	lada	156,561	476	329	6.30%	авг.19
4	zakon	132,098	3267	40	5.31%	авг.19
5	caravan	122,212	929	132	4.91%	авг.19
6	qostanay	95,824	1043	92	3.85%	авг.19
7	alau	94,975	491	193	3.82%	авг.19
8	informburo	79,118	1253	63	3.18%	авг.19
9	kstnews	66,572	338	197	2.68%	авг.19
10	mgorod	59,288	470	126	2.38%	авг.19
11	dknews	42,049	3318	13	1.69%	авг.19
12	uralskweek	39,33	194	203	1.58%	авг.19
13	kaz.tengrinews	37,674	725	52	1.51%	авг.19
14	elorda	33,342	444	75	1.34%	авг.19
15	ng	30,372	300	101	1.22%	авг.19
16	inbusiness	22,87	409	56	0.92%	авг.19
17	today	19,774	1089	18	0.80%	авг.19
18	tumba	13,772	320	43	0.55%	авг.19
19	ia-centr	12,839	159	81	0.52%	авг.19
20	forbes	11,003	538	20	0.44%	авг.19
21	toppress	10,791	230	47	0.43%	авг.19
22	ru.sputniknews	9830	1170	8	0.40%	авг.19
23	baribar	5033	82	61	0.20%	авг.19
24	mix.tn	3826	74	52	0.15%	авг.19
25	kp	3594	758	5	0.14%	авг.19
1	tengrinews	1,458,235	3017	483	57.56%	сен.19
2	lada	146,552	441	332	5.79%	сен.19
3	zakon	145,378	3597	40	5.74%	сен.19
4	caravan	134,557	769	175	5.31%	сен.19
5	qostanay	119,483	1126	106	4.72%	сен.19
6	alau	90,861	457	199	3.59%	сен.19
7	informburo	81,526	1198	68	3.22%	сен.19
8	mgorod	61,077	445	137	2.41%	сен.19
9	kstnews	52,915	339	156	2.09%	сен.19
10	kaz.tengrinews	50,103	734	68	1.98%	сен.19

11	ng	32,405	304	107	1.28%	сен.19
12	today	29,257	1123	26	1.15%	сен.19
13	uralskweek	24,186	160	151	0.95%	сен.19
14	dknews	20,667	2384	9	0.82%	сен.19
15	tumba	13,932	310	45	0.55%	сен.19
16	forbes	10,42	512	20	0.41%	сен.19
17	ia-centr	10,077	143	70	0.40%	сен.19
18	ru.sputniknews	10,001	1234	8	0.39%	сен.19
19	toppress	8516	149	57	0.34%	сен.19
20	elorda	6138	457	13	0.24%	сен.19
21	mix.tn	4213	78	54	0.17%	сен.19
22	kp	4007	759	5	0.16%	сен.19
23	bnews	3559	243	15	0.14%	сен.19
24	online.zakon	2015	342	6	0.08%	сен.19
25	knews	1932	172	11	0.08%	сен.19
1	tengrinews	1,426,898	3420	417	57.47%	окт.19
2	zakon	140,085	3769	37	5.64%	окт.19
3	caravan	128,938	796	162	5.19%	окт.19
4	lada	118,875	451	264	4.79%	окт.19
5	forbes	98,488	681	145	3.97%	окт.19
6	mgorod	93,006	475	196	3.75%	окт.19
7	qostanay	71,788	889	81	2.89%	окт.19
8	alau	68,059	492	138	2.74%	окт.19
9	informburo	63,414	1202	53	2.55%	окт.19
10	uralskweek	39,796	192	207	1.60%	окт.19
11	kstnews	38,317	315	122	1.54%	окт.19
12	kaz.tengrinews	30,453	756	40	1.23%	окт.19
13	today	23,913	1244	19	0.96%	окт.19
14	dknews	23,293	2258	10	0.94%	окт.19
15	baigenews	23,202	278	83	0.93%	окт.19
16	ng	22,557	293	77	0.91%	окт.19
17	tumba	10,264	288	36	0.41%	окт.19
18	ru.sputniknews	9925	1227	8	0.40%	окт.19
19	ia-centr	8157	140	58	0.33%	окт.19
20	toppress	7152	154	46	0.29%	окт.19
21	zirki	5920	1	5920	0.24%	окт.19
22	mix.tn	3946	77	51	0.16%	окт.19
23	kp	3897	821	5	0.16%	окт.19
24	elorda	3738	434	9	0.15%	окт.19
25	mtrk	2377	190	13	0.10%	окт.19
1	tengrinews	1,217,856	3032	402	50.82%	ноя.19
2	lada	133,927	497	269	5.59%	ноя.19
3	forbes	127,276	667	191	5.31%	ноя.19

4	zakon	107,006	3665	29	4.47%	ноя.19
5	mgorod	100,886	457	221	4.21%	ноя.19
6	baigenews	99,749	1037	96	4.16%	ноя.19
7	caravan	95,825	714	134	4.00%	ноя.19
8	qostanay	79,93	925	86	3.34%	ноя.19
9	alau	75,639	461	164	3.16%	ноя.19
10	informburo	61,311	1231	50	2.56%	ноя.19
11	uralskweek	60,726	206	295	2.53%	ноя.19
12	kstnews	56,464	340	166	2.36%	ноя.19
13	today	32,998	1235	27	1.38%	ноя.19
14	kaz.tengrinews	31,94	732	44	1.33%	ноя.19
15	ng	24,2	300	81	1.01%	ноя.19
16	baribar	19,498	238	82	0.81%	ноя.19
17	dknews	14,112	2171	7	0.59%	ноя.19
18	ru.sputniknews	10,26	1203	9	0.43%	ноя.19
19	tumba	9092	243	37	0.38%	ноя.19
20	ia-centr	7658	132	58	0.32%	ноя.19
21	toppress	4581	127	36	0.19%	ноя.19
22	kp	3919	760	5	0.16%	ноя.19
23	mix.tn	2512	56	45	0.10%	ноя.19
24	altyn-orda	2412	259	9	0.10%	ноя.19
25	mtrk	2089	202	10	0.09%	ноя.19
1	tengrinews	1,304,803	2940	444	53.89%	дек.19
2	lada	127,425	514	248	5.26%	дек.19
3	zakon	116,296	3276	35	4.80%	дек.19
4	forbes	113,629	644	176	4.69%	дек.19
5	informburo	95,749	1107	86	3.95%	дек.19
6	mgorod	85,828	419	205	3.54%	дек.19
7	caravan	79,733	648	123	3.29%	дек.19
8	qostanay	68,924	870	79	2.85%	дек.19
9	alau	67,725	439	154	2.80%	дек.19
10	baigenews	59,755	875	68	2.47%	дек.19
11	kstnews	45,304	295	154	1.87%	дек.19
12	today	42,339	1088	39	1.75%	дек.19
13	kaz.tengrinews	41,604	783	53	1.72%	дек.19
14	uralskweek	40,304	210	192	1.66%	дек.19
15	vrk	36,91	316	117	1.52%	дек.19
16	ng	20,212	259	78	0.83%	дек.19
17	dknews	16,758	1767	9	0.69%	дек.19
18	ia-centr	9228	154	60	0.38%	дек.19
19	tumba	9081	244	37	0.38%	дек.19
20	ru.sputniknews	8968	1075	8	0.37%	дек.19
21	toppress	5798	94	62	0.24%	дек.19

Приложение Б

Алгоритм на основе шинглов для сравнения текстов

Перед сравнением текст проходит минимальные чистки и изменения:

- убираются html вставки такие как <strong>

- символы преобразуются в нижний регистр

- убираются запятые, точки, апострофы, знаки переноса строки, двойные пробелы, слешы.


```
<form method="post" action="
    <?=$_SERVER['PHP_SELF']?>">
    <strong>Оригинальный текст</strong>:
    <br />
    <textarea id="text1" name="text1" style="width: 100%; height: 200px;">
        <?=isset($_POST['text1']) ? stripslashes(htmlspecialchars($_POST['text1'])) : "?>
    </textarea>
    <br />
    <strong>Переделанная (рерайт) копия</strong>:
    <br />
    <textarea id="text2" name="text2" style="width: 100%; height: 200px;">
        <?=isset($_POST['text2']) ? stripslashes(htmlspecialchars($_POST['text2'])) : "?>
    </textarea>
    <br />
    <br />
    <input type="submit" value="Проверить" style="display: block; margin: 0 auto; font-weight: bold; width: 50%;" />
</form>
<p>
    <?php
function get_shingle($text,$n=3) {
    $shingles = array();
    $text = clean_text($text);
    $elements = explode(" ",$text);
    for ($i=0;$i<(count($elements)-$n+1);$i++) {
        $shingle = '';
        for ($j=0;$j<$n;$j++){
```

```
                $shingle .= mb_strtolower(trim($elements[$i+$j]), 'UTF-8')." ";
            }
            if(strlen(trim($shingle)))
                $shingles[$i] = trim($shingle, ' -');
        }
        return $shingles;
}
function clean_text($text) {
        $new_text = eregi_replace("[\,|\.|\'|\"|\\|\/]","",$text);
        $new_text = eregi_replace("[\n|\t]"," ",$new_text);
        $new_text = preg_replace('/(\s\s+)/', ' ', trim($new_text));
        return $new_text;
}
function check_it($first, $second) {
        if (!$first || !$second) {
            echo "Отсутствуют оба или один из текстов!";
            return 0;
        }
        if (strlen($first)>200000 || strlen($second)>200000) {
            echo "Длина обоих или одного из текстов превысила допустимую!";
            return 0;
        }
        for ($i=1;$i
                        <5;$i++) {
            $first_shingles = array_unique(get_shingle($first,$i));
            $second_shingles = array_unique(get_shingle($second,$i));
            if(count($first_shingles) < $i-1 || count($second_shingles) < $i-1) {
                echo "Количество слов в тексте меньше чем длинна шинглы
                        <br />";
                continue;
            }
            $intersect = array_intersect($first_shingles,$second_shingles);
            $merge                                                          =
array_unique(array_merge($first_shingles,$second_shingles));
            $diff = (count($intersect)/count($merge))/0.01;

            echo "Количество слов в шингле - $i. Процент схожести -
".round($diff, 2)."%
                        <br />";
        }
}
```