УДК 316.77
DOI: 10.25629/HC.2023.01.14

ТЕХНОЛОГИИ АНАЛИЗА СОЦИАЛЬНЫХ СЕТЕЙ С ЦЕЛЬЮ ВЫЯВЛЕНИЯ СОЦИАЛЬНЫХ ТРЕНДОВ

Галаганова С.Г., Турусина Т.В.
Московский государственный технический университет им. Н.Э. Баумана

Аннотация. Одним из условий эффективного управления современным социумом и его подсистемами является своевременное и точное определение социальных трендов – устойчивых тенденций и доминирующих направлений социальной динамики. В условиях цифровой трансформации общества главной площадкой социальных коммуникаций и, следовательно, сбора эмпирических данных стали социальные сети. Однако объём информации в социальных сетях намного превосходит возможности её обработки обычными способами, создавая потребность в алгоритмизации и автоматизации данного процесса. Авторами статьи предпринята попытка разработки аналитического алгоритма для последующего создания соответствующего программного продукта. В статье рассмотрены законы ценности социальной сети Д. Сарноффа, Р. Меткалфа, Д. Рида и Дж. Ципфа, проанализированы свойства социальных сетей, представленные в виде социальных графов, определены метрики, описывающие силу влияния социально-психологических эффектов на пользователя сети, выявлена специфика медиатекстов в социальных сетях, выведены корреляции между социальными эффектами и формированием социальных медиатрендов. Результаты работы могут послужить основой для разработки программного продукта для автоматизированной обработки корпуса медиатекстов в социальных сетях.

Ключевые слова. Социальная сеть, социальный тренд, трендвотчинг, семантический анализ, закон ценности социальной сети, социальный граф, сильные связи, слабые связи, корпус текстов.

Введение

Социальный тренд – доминирующее направление социального развития, устойчивая тенденция изменения общественного процесса или явления. Своевременное и точное выявление социальных трендов является необходимым условием научного управления социумом и его подсистемами, адекватного целеполагания и аксиологического ориентирования. Для выявления социальных трендов сегодня используются самые разные методы научного познания, однако начальным звеном по-прежнему остаётся эмпирическое наблюдение и последующий анализ эмпирических данных. Это, в свою очередь, обусловливает актуальность грамотного выбора социального пространства (площадки) для сбора необходимой информации [13].

В условиях цифровой трансформации социума, затрагивающей почти все сферы общественной жизни, таким пространством стали социальные сети [1; 17; 24]. Пользователи Интернета проводят в них по несколько часов в сутки, общаясь, читая новости, организуя свой досуг, формируя сообщества по интересам. Подтвердив прогнозы американского социолога Мануэля Кастельса [10], социальные сети превратились в «виртуальное зеркало» информационного социума и эффективный инструмент его формирования [7]. Несмотря на выводы ряда исследователей о конвергентном характере современной медиакультуры [9], именно социальные сети являются сегодня главной площадкой социальной коммуникации, а распространяемые в них медиатексты – главной формой отражения массовых социальных ожиданий и притязаний, неиссякаемым источником эмпирических данных для отслеживания социальной динамики [3].

Это, в частности, нашло отражение в перемещении в сетевое пространство так называемого трендвотчинга (англ. trendwatching) – практики регулярного отслеживания маркетологами по-

требительских трендов. Social media marketing (SMM), т.е. маркетинг в социальных сетях, сегодня является важнейшей составной частью маркетинговой и коммуникационной стратегии любой значительной компании. Он представляет собой систему мероприятий по использованию социальных медиа в качестве каналов для продвижения бренда и решения самых различных бизнес-задач – от оценки и прогнозирования бизнес-факторов (спроса, предложения, моды) до создания инновационных продуктов, услуг и коммуникаций [2].

Однако объём информации в социальных сетях делает невозможной её обработку и систематизацию традиционными методами – необходима алгоритмизация и автоматизация данного процесса. Этим обусловлена актуальность и практическая значимость данного исследования.

Цель и задачи исследования

Авторами статьи предпринята попытка разработки аналитического алгоритма для последующего создания программного продукта, позволяющего автоматизировать применение методов семантического анализа текстовых данных в социальных сетях с целью выявления социальных трендов.

Для реализации данной цели решались следующие задачи:

- выявить специфику медиатекстов в социальных сетях;

- исследовать и сравнить используемые в настоящее время методы семантического анализа медиатекстов в социальных сетях;

- осуществить сбор и первичную обработку текстовых данных одной из социальных сетей;

- создать алгоритм семантического анализа текстовых данных в социальных сетях для выявления социальных трендов.

Будем считать, что рассматриваемая в работе сеть располагает следующим набором возможностей: создание персонального профиля, добавление друзей, объединение в сообщества с другими, создание и продвижение контента.

В качестве социального тренда применительно к виртуальной сети авторами рассматривалась любая устойчивая смысловая общественная тенденция, проявляющаяся в массиве пользовательского контента.

Методология исследования

Исследователи социальных сетей выделяют феномен *«силы слабых связей»*. Различают «сильные» и «слабые» социальные связи, критерием разделения выступает частота и длительность контактов. Примером «сильных» связей являются родственные и дружеские связи, «слабых» – формальные контакты (коллеги, соседи и т.д.). Американский социолог М. Грановеттер определил, что внутри социальных сетей по «слабым» связям информация быстрее и шире распространяется, нежели по «сильным», следовательно, «слабые» связи имеют большее значение [6]. По мнению учёного, «слабые» связи крайне необходимы для расширения возможностей взаимодействия пользователей друг с другом и с сообществом, тогда как в результате «сильных» связей возникает лишь локальный контакт. Через «сильные» связи люди делятся ограниченным объёмом данных, поэтому эти связи информационно избыточны, а их взаимная польза незначительна.

Влияние «силы слабых связей» для сети с определённым количеством участников можно рассчитать математически. Данная метрика называется *ценностью социальной сети* [6]. Существует несколько различных подходов к определению и вычислению ценности социальной сети, которые можно представить в виде таблицы 1.

Таблица 1 – Подходы к определению и вычислению ценности социальной сети

Автор подхода	**Смысл ценности сети**	**Формула зависимости ценности сети от количества её участников n**
Дэвид Сарнофф	Доступность пользователей друг для друга	n
Роберт Меткалф	Количество возможных контактов	$n(n-1)$
Дэвид Рид	Потенциал создания сообществ	$2^n - n - 1$
Эндрю Одлыжко («закон Ципфа»)	Качество связей	$n * ln\, n$

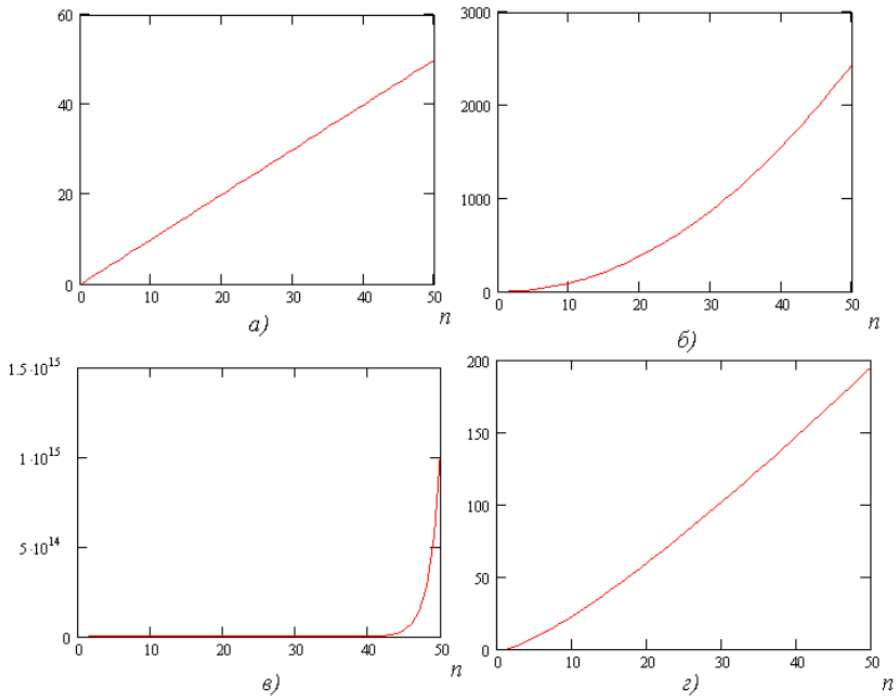

Рисунок 1 – Иллюстрация изменения ценности социальной сети при увеличении числа её участников от 0 до 50 в приложении к законам: а) Сарноффа, б) Меткалфа, в) Рида, г) Ципфа

Поскольку ценность сети обусловлена количеством связей, важной задачей становится измерение информационного взаимовлияния пользователей. Это можно сделать при помощи социальных графов [4]. Социальный граф – это граф, узлы которого представлены социальными объектами, такими, как пользовательские профили с различными атрибутами, сообщества, элементы медиаконтента и т.д., а рёбра – социальными связями между ними. Графы являются удобной и наглядной формой представления информации для последующего анализа (рис. 2).

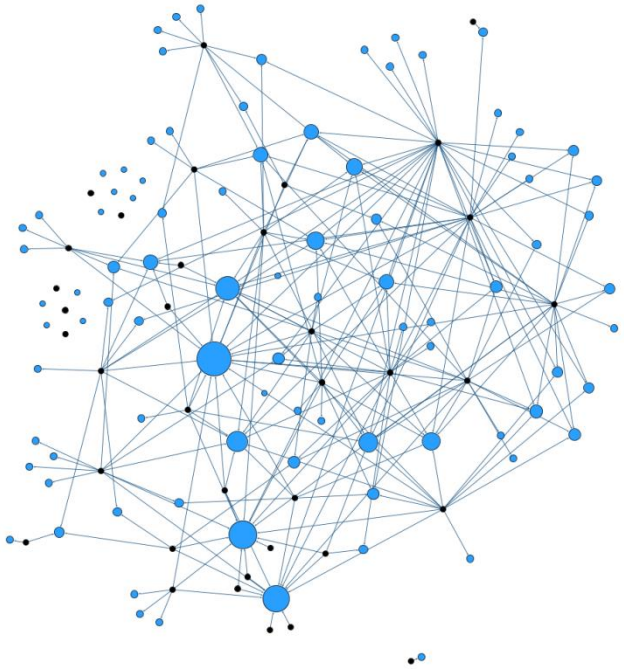

Рисунок 2 – Пример социального графа. Наиболее крупными точками отмечены узлы с наибольшим распределением на степенях вершин

Социальные графы характеризуются следующим набором свойств.

Одна большая общая компонента связности. Каждый пользователь связан с любым другим пользователем из сети. Из этого следует, что граф является связным, а между любыми двумя вершинами существует путь. Конечно, возможно, что в социальной сети зарегистрированы пользователи, никогда не взаимодействующие с остальными (не входящие в общую компоненту связности), данные пользователи не получают и не распространяют информацию, следовательно, могут быть исключены из исследования.

Распределение на степенях вершин. Степень вершины – это количество рёбер графа, инцидентных данной вершине. Её распределение есть распределение вероятности, что степень вершины будет соответствовать некой случайной величине. Социальные сети относятся к так называемым безмасштабным сетям.

Среднее расстояние. Расстояние между вершинами в графе – число рёбер в кратчайшем пути, соединяющим эти вершины. В социальных сетях пути являются каналами передачи информации. Причём путей между двумя случайными вершинами может быть множество, их количество зависит от степеней вершин. Обычно, среднее расстояние между акторами социальных сетей небольшое, поэтому информация быстро передаётся внутри сети.

Коэффициент кластеризации. Кластер – это выборка наиболее социально связанных пользователей, что в данном случае означает совпадение набора подписок (когда группируются потребители одинакового контента). Коэффициент кластеризации показывает, насколько склонны пользователи к образованию сообществ: вершины, входящие в одну группу, соединены между собой гораздо плотнее, чем со всем остальным графом.

Структура сообществ. Сообщества выделяются из графа на основе кластеров. Структура сообществ, в том числе их связность, позволяет выявить интересы и настроения пользователей.

В результате анализа графов можно определить, в каких сообществах (группах по интересам) состоит пользователь, установить круг его общения, личные предпочтения, его авторитет в сети. Эта информация, в свою очередь, облегчит идентификацию пользователей, социальный поиск, выявление неявных связей.

Ещё одной важной задачей, которую позволяет решить анализ социальных графов, является выявление логики организации социальной сети, т.е. признаков, в соответствии с которыми организованы её объекты (пользователи и их сообщества). Для этого, более глубинного, поиска корреляций, требуется ввод дополнительных (частично не упомянутых выше) параметров. В зависимости от искомых характеристик, исследователь может опираться в данном случае на следующие группы метрик:

1) *Метрики, отражающие природу взаимоотношений между социальными объектами.*

- гомофилия, т.е. свойство образовывать связи с схожими узлами (люди с одинаковыми увлечениями, общим районом проживания, местом учёбы или работы наверняка будут связаны между собой);

- множественность, т.е. свойство образовывать связи разных типов между объектами (люди одновременно родственники и коллеги). С помощью данной метрики можно отслеживать тип связей (“слабые” или “сильные”);

- взаимность связей;

- сетевая закрытость, т.е. мера взаимности связей между узлами, с которыми связан пользователь (все друзья пользователя связаны и с ним, и друг с другом). В ситуации при практически полной сетевой закрытости, у пользователя сформируется «информационный пузырь», человек начинает получать только информацию, соответствующую общеразделяемым интересам в компании;

- соседство, т.е. мера связи с географически близкими объектами (жители одного города).

2) Метрики, отражающие характер связей, как для отдельных социальных объектов, так и для графа в целом.

- мост, т.е. объект, через который проходят единственная связь между кластерами или пользователями;

- структурные дыры, т.е. полное отсутствие связей между двумя частями сети;

- центральность [16].

Последнее понятие имеет особое значение для анализа социальных сетей, поскольку позволяет выявлять узлы, играющие особую роль в транслировании информации и поддержании целостности сети.

Вышеописанные метрики позволяют выявить сильные и слабые связи, найти влиятельных пользователей, определить плотность связей в сети.

Существуют следующие показатели центральности:

1. Центральность по степени

Центральным объектом является тот, который обладает максимальным количеством связей. Данная метрика показывает авторитетность узла, при условии, что связи объекта качественны, т.е. не являются результатом автоматической «накрутки» подписчиков.

В случае, если нам нужно сравнивать центральные по степени узлы в рамках разных сетей, для расчёта центральности применяется формула:

$$C_D\,(i)\ = \frac{\sum_{i \neq j}^{n} g_{ij}}{n-1},$$

Где $C_D\,(i)$ – степень центральности узла i,

n – число вершин в сети,

g_{ij} – связь между узлами i и j.

Существуют вариации центральности по степени: входящая и исходящая. Узел входящей центральности обладает наибольшим количеством входящих дуг. В социальных сетях таким узлом является блоггер - лидер по числу подписчиков, который является источником информации для множества других объектов.

Аналогично, исходящая центральность определяется числом исходящих связей, и указывает на пользователя с большим количеством друзей и подписок. Такой актор может передавать информацию в разнородные сообщества, пересылая посты одного знакомого другому (по «слабым связям»). Собеседники склонны больше доверять персонифицированному источнику, особенно если они контактировали с ним.

2. Центральность по близости

Центральным по близости является объект, который расположен наиболее близко к прочим узлам сети. Таким узлом может являться пользователь с большим количеством подписчиков, подписанный также на большое количество других пользователей. Данная метрика отражает коммуникативную эффективность актора и вычисляется по формуле:

$$C_c(i) = \frac{1}{\sum_{j=1}^{n} d(i,j)} \times \frac{1}{n-1},$$

Где $C_c(i)$ плотность центральности узла i,

$d(i,j)$ – расстояние между вершинами i и j,

n – число вершин в сети.

Центральный по близости пользователь постоянно находится в информационном потоке, поглощая, перерабатывая и распространяя по сети новые данные, таким образом его деятельность определяет скорость передачи новостей по сети.

3. Центральность по посредничеству

Центральность по посредничеству означает, что через данный узел проходит максимальное число кратчайших путей между всеми узлами сети. Такой пользователь может являться мостом и контролировать связи между разными сообществами. Данная метрика указывает на неотъемлемого участника всех процессов сети. В случае, если пути, исключающие данного пользователя, нестабильны и малоэффективны, центральный актор сможет получить монополию на фильтрацию или редактирование информации.

Формально, центральность по посредничеству определяется следующим образом:

$$C_B(i) = 2\frac{\sum_{j \neq k} g_{jk}(i)/g_{jk}}{(n-1)(n-2)},$$

где $C_B(i)$ – центральность как посредничество узла i,

$g_{jk}(i)$ – число самых коротких путей, соединяющих j и k и проходящих через вершину i ,

g_{jk} – общее количество коротких путей, соединяющих j и k,

n – число вершин в сети.

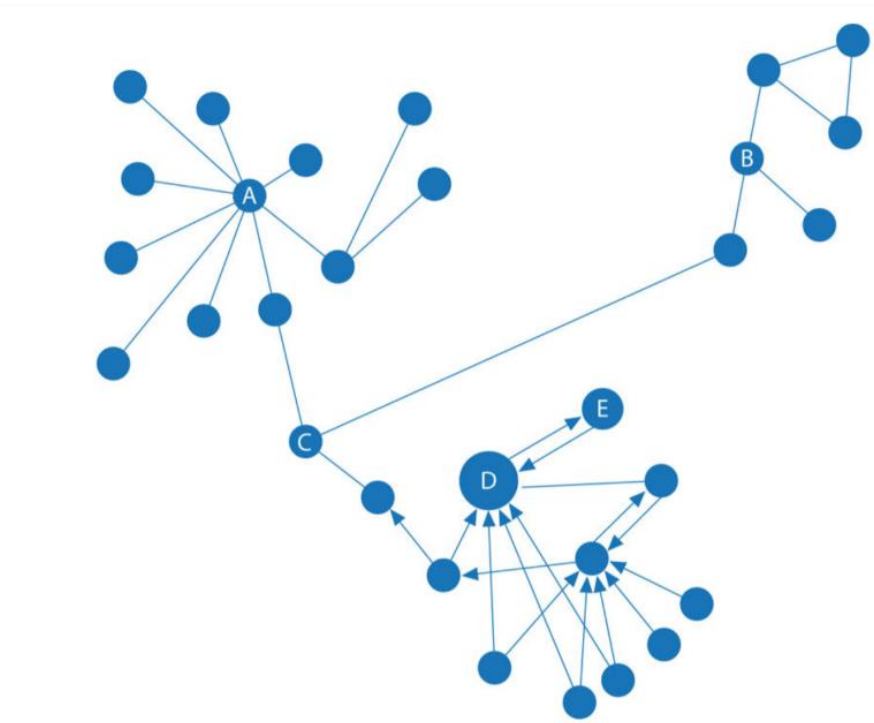

Рисунок 3 – Социальный граф, на котором выделены: а) центральность по степени в рамках своего кластера(A), б) центральность по посредничеству в рамках своего кластера(B), в) мост(C), г) центральность по близости в рамках своего кластера(D)

3) *Метрики, отображающие характеристики социального графа, поделённого на сегменты, которые имеют отличительные особенности.* Они используются для обнаружения особо плотных по связям районов в пределах единой сети. К ним относятся клика – группа, в которой все пользователи имеют прямые связи друг к другу (группа, в которой такие связи не обязательны, называется социальным кругом) и коэффициент кластеризации.

Методика и инструментарий исследования

Можно выделить следующие основные социально-психологические характеристики социальных сетей, оказывающие влияние на формирование социальных трендов [8]:

1) Целенаправленное активное поведение пользователей.

Люди используют социальные сети, чтобы удовлетворить собственные потребности в общении, получении новостей, комфорте и отдыхе. Реализация данных потребностей определяет эмоциональную вовлеченность в извлекаемый из сети контент.

2) Формирование групп.

Акторы сети обычно объединяются в подгруппы на основе общих интересов, потребностей, близости мнений. На поведение пользователей, состоящих в сообществе, воздействуют психологические групповые эффекты: у людей формируется чувство групповой солидарности, идентичности. Таким образом, самооценка этих людей начинает зависеть от статуса сообщества (например, числа единомышленников). Также, за счёт эффекта подражания, привычки и взгляды личности корректируются под влиянием группы.

3) Подверженность пользователей влиянию других акторов социальной сети и изменение личных мнений под их влиянием.

Эффект конформизма заставляет людей подстраивать свои воззрения под мнения группы. Степень проявления этого эффекта зависит от численности группы, статуса «оценщиков» (тяжелее переносится критика со стороны лидеров мнений), сплочённости группы (чем больше у человека союзников, тем сложнее изменить его взгляды), психологической среды (в группах с агрессивной риторикой процент конформизма выше).

4) Наличие специфических социальных норм в виртуальном пространстве.

5) Существование внешних факторов влияния (реклама, фейки, спам) и, соответственно, внешних агентов, стремящихся контролировать или навязывать общественные тенденции.

6) Наличие определённых этапов динамики мнений членов социальной сети:

Информация распределяется по сети следующим образом:

- один из членов сообщества высказывает мнение;

- идея распространяется среди других участников группы, причём исходный тезис может как изменяться (например, если пользователи дополнили или мисинтерпретировали его), так и сохраняться (например, благодаря репостам оригинального поста);

- устанавливается единая трактовка идеи в сообществе, обычно это происходит после высказывания мнения лидеров группы;

- идея покидает пределы группы: пользователи начинают ретранслировать мнение в других виртуальных сообществах, высказываться в комментариях или на личных страницах.

7) воздействие структурных свойств социальных сетей на динамику мнений:

- существование взаимного косвенного влияния акторов сети по цепочке социальных контактов («слабым связям»): чем меньше у пользователей общих друзей и сообществ, тем меньше вероятность солидаризации с мнением друг друга;

- чем больше у пользователя связей, тем, с одной стороны, больше у него возможностей через своё окружение повлиять на всю сеть, а с другой – большая подверженность чужому влиянию;

- в рамках одного кластера выстраивается «сильная связь», а главное действующее лицо кластера выполняет функцию ретранслятора;

- каждое сообщество может быть рассмотрено как кластер людей с определённым спектром мнений;

- локальная промежуточность определяет актора моста, который может получать информацию из разных кластеров одновременно (данный актор не может выступать лидером мнений, однако имеет «слабые связи» с множеством групп).

Медиатексты в социальных сетях существенно отличаются от публицистических произведений или репортажей СМИ, поскольку создаются рядовыми пользователями. Поэтому анализ контента социальных сетей предполагает предварительное выявление его специфики, которую можно свести к следующим основным характеристикам:

1. Тексты часто написаны нелитературным языком, в них могут содержаться иностранные слова, сленг, обсценная лексика. Тексты могут быть грамматически неверны, содержать орфографические ошибки. Причём всё это может быть включено в текст намеренно, для трансляции неких мета-смыслов.

2. Пользователи реагируют на текст посредством лайков, репостов и комментариев. Причём их отклик подвергается количественной оценке.

3. Текст и реакции быстро устаревают в связи с высокой скоростью обновления данных в Интернете.

4. Тексты могут отличаться так называемым нестабильным качеством: это может быть спам, информационный шум.

5. Тексты содержат эмоциональную окраску, выражающуюся в тональности сообщений.

При проведении аналитической работы на больших массивах медиатекстов следует также учитывать ограничения, обусловленные спецификой платформы:

1. Неструктурированность получаемой информации: массив исходных данных собирается с использованием программных интерфейсов социальных сетей(API), функционал которых часто ограничен. API выполняют функцию порта, через который внешняя программа получает доступ к данным сети. Происходит процесс сбора информации путём отправки запросов, предусмотренных разработчиками интерфейса. С помощью одного запроса можно получить список подписчиков сообщества или пользователя, однако невозможно обнаружить акторов с близкими интересами и общими друзьями. Таким образом API защищает базы данных сети от утечки, при этом предоставляя функционал для сбора информации и интеграции. Проблемой, в рамках решаемой задачи, становится необходимость структурировать данные и обнаруживать взаимосвязи.

2. Блокировки доступа: пользователи имеют право скрывать свои личные данные и ограничивать доступ к личным страницам.

3. Огромный объём исходной информации: каждую минуту публикуется большое количество медиатекстов даже в кругу инфлюенсеров, и обрабатывать каждый пост невозможно в силу ограничений вычислительной мощности. Решением данной проблемы является построение репрезентативной выборки.

Что касается прикладных способов извлечения смыслового содержания из сетевых сообщений, то все их можно разделить на две основные группы: методы *лингвистического* анализа, основанные на извлечении смысла текста по его семантической структуре, и методы *статистического* анализа, основанные на извлечении смысла по частотному распределению слов в тексте [11].

Для выявления социальных трендов имеет смысл использовать методы обеих категорий в определённом сочетании. В настоящее время в практике исследования пользовательского контента социальных сетей наиболее активно используются следующие варианты семантического анализа [12]:

1) Метод *частотно-семантического анализа (ЧСА)*.

В сущности, метод сводится к подсчёту частотности слов в тексте. Во избежание искажения данных, метод рассчитывает встречаемость сем, то есть минимальных носителей содержания.

Обычно при использовании данного метода семами считаются только существительные. Подобное ограничение обосновывается некоторыми лингвистическими концепциями о членении предложений, согласно которым тема текста есть его субъект (агенс). Агенсами могут являться существительные, местоимения и субстантивацинированные прилагательные (имена прилагательные, перешедшие в разряд существительных в рамках контекста). При анализе медиатекстов, написанных в разговорном стиле, невозможно однозначно определить предмет, на который указывает местоимение, поэтому от их подсчёта отказались.

Метод частотно-семантического анализа использует словарь для выделения сем в тексте, теоретически субстантивацинированные прилагательные, являющиеся частью Интернет-сленга, также могут быть в него занесены.

2) *Стемминг* (от англ. stem – «корень», «основа») – процесс отсечение от слова окончаний и суффиксов для нахождения стеммы – основы слова, от которого образуются все его грамматические формы. Одним из распространённых алгоритмов стемминга является *Стеммер Портера,* подробнее описанный в таблице ниже. Стеммеры может работать только с языками, которые реализуют словоизменение через аффиксы. Преимущество стеммеров в отсутствии необходимости использования словаря [5]. Вместе с тем, стеммеры имеют ряд недостатков:

- алгоритмы могут ошибочно обрабатывать и имена собственные, тем самым искажая результат исследования (например, «Великие Луки»);

- в русском языке суффиксы часто несут смысловую нагрузку, при сокращении слова до основы возможна потеря мета-смыслов – например, оценочного суждения автора текста (например, «человечище», «человечишка»).

3) *Латентно-семантический анализ (ЛСА)* – один из методов обработки информации на естественном языке (Natural Language Processing). Идея данного метода состоит в нахождении ассоциативно-семантических связей, то есть скрытых связей между словами, которые отображают контекстную или ассоциативную близость слов в коллекции текстов. Осуществляется это методами линейной алгебры, за счёт «сжатия» исходного текстового набора данных.

ЛСА получает в качестве входных данных матрицу термы-на-документы. Она формируется следующим образом:

- собирается массив документов, в нашем случае медиатекстов;

- из документов извлекаются и заносятся в словарь все различные термы (слова или любые другие последовательности символов);

- вычисляется вес каждого терма в соответствующем документе, либо как частота слова, либо как произведение количества вхождений слова в документ на переменную, обратнопропорциональную количеству документов, в которых встречается терм;

- создаётся матрица, строки которой отражают документы, столбцы - термы, а значения в матрице соответствуют весам термов.

Полученная матрица термы-на-документы – значительной размерности, следовательно, выявлять в ней зависимости проблематично. Для решения этой проблемы поводится сингулярное разложение матрицы, которое позволяет выделить ключевые составляющие матрицы, игнорируя шумы. С помощью сингулярного разложения любую матрицу можно разложить в виде произведения ортогональных матриц, комбинация которых является достаточно точным приближением к исходной матрице. Таким образом можно выявить наиболее значимые для описания исходных данных элементы, и уже анализируя их, определять корреляции терм.

4) *Синтаксико-семантический анализ.*

Методически, данный вид анализа сводится к выделению семантических отношений – смысловых связей между лексическими единицами, которые позволяют определить структуру текста и определить синтаксическую функцию слов.

Определяя семантические отношения, можно выделить позицию слова: главное оно или зависимое, а также определить группу, к которой принадлежит данная семантическая связь (ПРИЗНАК, ДЕЙСТВИЕ, МЕСТО и т.д.).

Для автоматизации определения семантических связей используются семантические шаблоны, состоящие из следующих элементов:

- «Словаря», содержащего последовательности слов, для которых уже определены семантические отношения;

- Названий семантических отношений, описанных в «словаре»;

- Численно обозначенных позиций слов в последовательности из «словаря», элементы которой должны быть добавлены в очередь с приоритетом;

- Приоритета группы семантических отношений, выражаемый численно.

В соответствии с очередью приоритетов слова будут последовательно удаляться из анализируемого предложения, до того момента, пока не будет выделено главное слово или словосочетание.

В обобщённо-схематизированном виде определение тематики текста на основе вышеуказанных методов семантического анализа можно представить в виде таблицы 2.

Таблица 2 – Определение тематики текста на основе методов семантического анализа

Метод семантического анализа	Алгоритм	Способ определения тематики текста
Частотно-семантический анализ	1. Каждое слово текста сравнивается со словарём; 2. Если слово есть в словаре, оно заносится в массив; 3. Подсчитывается число вхождений для каждого слова в массиве.	Слова с самым большим числом вхождений будут темой текста.
Стеммер Портера	1.Стемминг слов: a) Удаляются окончания деепричастий или окончания «ся» («сь»); b) Удаляются окончания прилагательных, глаголов и существительных; c) Удаляются окончания в оставшихся словах (по списку) «и», «ост», «ость», «ейш», «ейше», «ь»; d) Удаляется одна буква в окончании «нн». 2. Формируется массив основ слов, которые сравниваются по числу вхождений.	Стемма(основа) с наибольшим числом вхождений будет являться тематикой текста.
Латентно-семантический анализ	1. Составляется матрица термы-на-документы: $A=USV^T$, где матрицы U и V – ортогональные, а S – диагональная матрица, значения на диагонали которой называются сингулярными значениями матрицы A. Матрица V транспонирована. 2. Если при таком разложении в матрице S оставить только k наибольших сингулярных значений, а в матрицах U и V – только соответствующие этим значениям столбцы, то $\hat{A} \approx A=USV^T$. 3. $\hat{A}$ отображает основную структуру различных зависимостей, присутствующих в исходной матрице, где k - факторы связи терм.	Тематика текста определяется как корреляция между термами и документами из разложения.
Синтаксико-семантический анализ	1. Формируются базовые шаблоны, в том числе устанавливаются приоритеты для каждого вида семантических связей. 2. Производится составление очереди (массива шаблонов), позиция в которой основана на приоритете связи; 3. В соответствии с очередью зависимые слова удаляются из предложения. 4. Из каждого предложения, исходя из очереди с приоритетом, удалением определяется слово с наибольшим числом зависимостей и считается число его вхождений в текст.	Тематика текста определяется как наибольшая частота вхождений некого главного слова.

Результаты исследования и их обсуждение

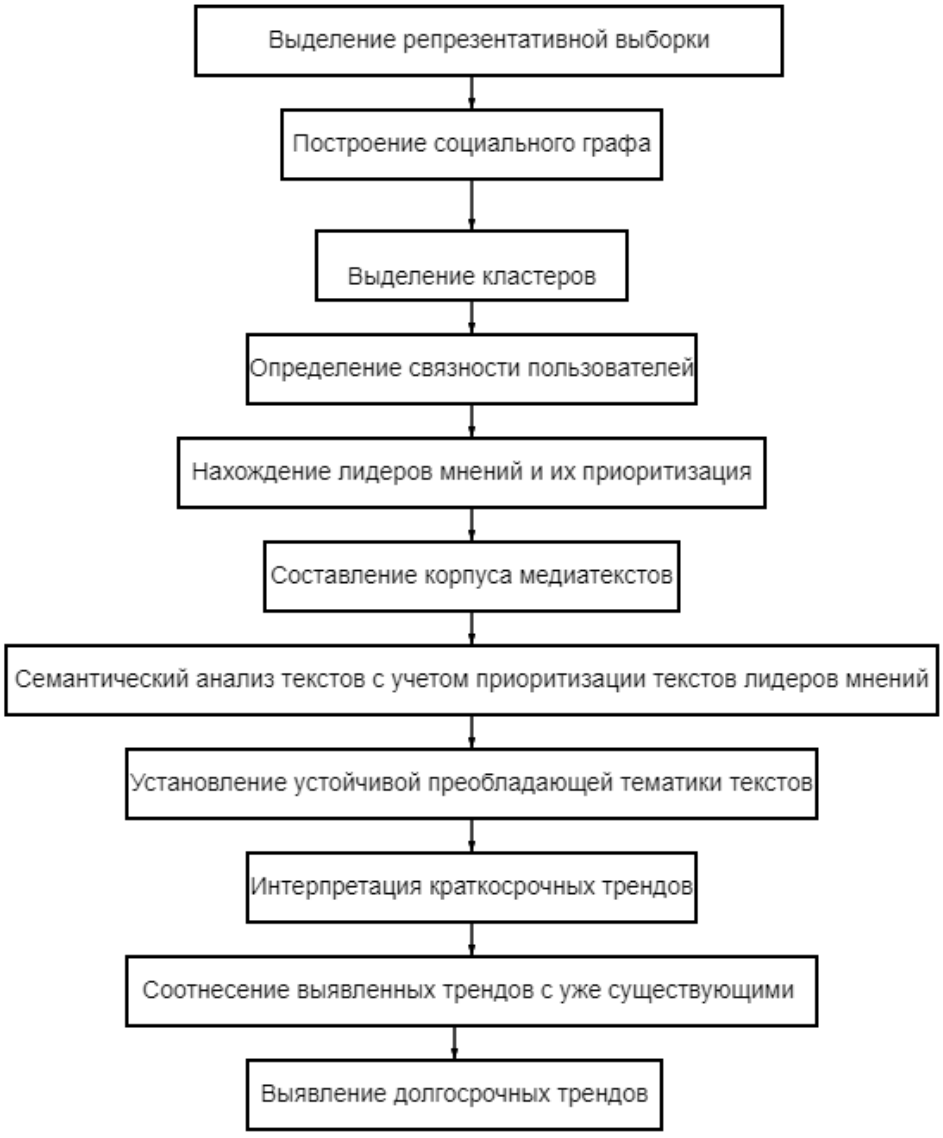

Рисунок 4 – Схема алгоритма выявления социальных трендов в виртуальных социальных сетях

1. Выделение репрезентативной выборки

Работа с выборкой в данном случае необходима, поскольку в силу ограниченности вычислительной мощности отслеживание постов всех пользователей невозможно. Репрезентативная выборка – это такая выборка из генеральной совокупности, в которой все основные признаки исходной совокупности представлены приблизительно в той же пропорции или с той же частотой.

2. Построение социального графа

Алгоритм построения социального графа включает построение матрицы, отражающей связи между членами выборки и сообществами, в которых они состоят; создание двудольного

графа, соответствующего данной матрице и присвоение веса каждой вершине в соответствии с количеством связей, проходящих через узел.

3. Выделение кластеров

Существует несколько различных алгоритмов, однако наиболее распостранённым является K-means (k-средних).

Основной тип задач, решаемых алгоритмом k-средних, – построение кластеров, которые являются максимально возможно различны, то есть средние в кластере (для всех переменных) максимально возможно отличаются друг от друга. Также с помощью данного метода возможно строить предположения относительно числа кластеров и проверять их.

Алгоритм выглядит следующим образом: сначала случайным образом генерируются k центроидов (фактически, произвольных точек), рассчитывается евклидово расстояние от каждого элемента до центроидов, после каждый элемент присваивается «ближайшему» центроиду, формируются условные кластеры. Далее проводится перерасчёт центров кластеров и назначаются новые центроиды, объекты снова присваиваются ближайшим кластерным центрам. Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из двух условий: кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации.

2. Определение связности пользователей

Для этого определяется плотность – отношение количества связей к общему числу акторов сети. Плотность отражает скорость, с которой информация может распространяться по сети (чем больше связей в сети, тем выше скорость). Плотность возможно рассчитать по формуле:

$$\rho = \frac{2m}{n(n-1)}$$

где m – количество рёбер в графе; n – количество вершин.

3. Нахождение лидеров мнений и их приоритизация

В случае, если нужно обнаружить только одного лидера в сообществе, необходимо просто определить узел, центральный по степени. Если же необходимо рассмотреть степень влиятельности каждого из пользователей, то для этого n раз происходит определение степени вершины, где n – число акторов сети. Участники сети с наибольшим количеством связей становятся наиболее влиятельными, поэтому их тексты имеют приоритет. В случае если лидеры мнений не очевидны, можно указать узел, центральный по близости.

4. Составление корпуса медиатекстов

В ситуации постоянного обновления микротрендов, общественное мнение за анализируемый период может многократно меняться, поэтому для получения однозначных выводов входные данные следует фильтровать по времени. Если же требуется узконаправленное исследование, то фильтрацию контента можно осуществить ещё в процессе сбора данных. Так, если задачей исследователя является выявление социальных трендов в политической сфере, бессмысленно тратить ресурсы и включать в корпус текстов посты из хобби-сообществ, юмористические или рекламные.

Большинство онлайн-платформ сегодня ведут борьбу со спамом и постами ботов. Тем не менее, проблема искажения корпуса текстов информационным шумом существует. Часто спам-тексты пишут и распространяют реальные люди с целью рекламы. Эти пользователи не просто копируют одно и то же сообщение, но и могут успешно вписывать спам в контекст обсуждения. Это создаёт необходимость очищать выборку от ссылок на другие сообщества или внешние сайты, от фраз, призывающих к переходу на страницу автора или совершению покупок, от повторяющихся сообщений, от сообщений, в комментариях к которым другие пользователи обозначали автора как бота.

В социальных сетях пользователи могут выражать свою поддержку медиатексту лайками или размещать его на своей странице (репост). Данные метрики позволяют рассчитать количественную оценку, выдвигаемую пользователями посту: отклик = (количество лайков + количество репостов) количество просмотров. Полученные данные откликов можно сравнивать между собой в рамках одного кластера при условии, что за определённый промежуток времени количество членов сообщества не менялось в пределах некой погрешности.

Нетипично большое количество комментариев может свидетельствовать о яркой эмоциональной реакции пользователей на пост. Для того, чтобы узнать является их отклик позитивным или негативным, необходимо соотнести количество комментариев с количеством лайков. Если большинство пользователей, оставивших комментарии, оценило исходный медиатекст, можно предположить, что их отклик является положительным.

5. *Семантический анализ текстов с учётом их весов, т. е. статуса авторов*

Он предполагает расстановку весов текстов на основе авторитета авторов, «домножение» текстов на веса (текст копируется в выборку несколько раз), за счёт чего происходит приоритизация текстов лидеров мнений, и проведение семантического анализа одним из вышеописанных методов.

6. *Определение стабильно преобладающей тематики текстов, т. е. тренда*

Социальным трендом считаются не только процессы, но и устойчивые мнения, нормы ценностные ориентации, которые находят отражение в медиатекстах [18]. Преобладание определённой тематики означает, что люди проявляют интерес к данной проблеме. Ярко выраженная эмоциональная оценка определённой тематики позволяет не только определить настроения в обществе, но и выделить перспективы развития социального тренда, уровень конформизма в дискурсе, оценить возможности появления противоположного, «реакционного» тренда.

7. *Трактовка краткосрочных трендов и формирование общей картины*

Зачастую простого определения тональности текста недостаточно для определения пользовательского отношения, которое может выражаться не «напрямую», а через использование иронии и сарказма. Автоматизировать подобное достаточно проблематично, поэтому на данном этапе требуется участие оператора, который бы просмотрел несколько текстов, содержащих тренд, от наиболее влиятельных авторов. В этом случае оператор может подтвердить или опровергнуть выдвинутую гипотезу об эмоциональной оценке пользователей.

Полученные на основе п. 1 данного алгоритма тренды как факт существуют на момент формирования выборки медиатекстов, однако возникает вопрос: являются ли выявленные тренды чем-то принципиально новым? Для ответа на него тренды фиксируются в документах и группируются для составления полной картины. Составление наглядных схем выявленных трендов является крайне важной операцией ещё и в связи с большим количеством микро-трендов, распространяемых в сети. Соотнесение краткосрочных трендов по группам-кластерам позволяет системе в дальнейшем успешнее сравнивать существующие и выявленные тренды, а также позволяет аналитикам увидеть формы, в которых могут выражаться близкие по содержанию тенденции.

8. *Сопоставление недавно выявленных трендов с уже известными*

Далее требуется выяснить, является выявленный тренд новинкой или проявлением уже продолжительно существующей тенденции. Во втором случае потребуется установить, является ли данный показатель усилением или ослаблением общей тенденции.

9. *Выявление долгосрочных трендов*

Выводы

Пространство социальных сетей превращается сегодня в одну из главных сфер проявления социальной динамики, а сетевой медиаконтент – в базу данных социальной аналитики. Применение методов семантического анализа к исследованию корпуса сетевых медиатекстов позволяет оперативно выявлять стихийно формирующиеся, не вполне очевидные тенденции и формы их проявления, что, в свою очередь, является необходимым условием эффективного

социального управления. Однако объём сетевой информации исключает возможность её результативной обработки традиционными способами, требуя автоматизации данного процесса. Алгоритмизация аналитического инструментария лингвистической семантики становится одним из актуальных направлений современной науки. В этой связи разработанный авторами статьи алгоритм выявления социальных трендов имеет не только эвристическое, но и практическое значение в качестве основы для последующего создания соответствующих программных продуктов.

Библиография

1. Володенков С.В. Интернет-коммуникации в глобальном пространстве современного политического управления: навстречу цифровому обществу. М.: Проспект, 2021. 416 с.

2. Галаганова С.Г., Мартынова А.А. Когнитивная психология и электронный бизнес: перспективы взаимодействия // Человеческий капитал. № 3(159). 2022. С. 23-40.

3. Галаганова С.Г. Социокультурное измерение цифровой трансформации // Этносоциум и межнациональная культура. 2022. № 10(160). С. 27-33.

4. Гитис Л.Х. Статистическая классификация и кластерный анализ. М.: МГГУ, 2003. 157 с.

5. Головко Н.В. Оценка семантического потенциала текстов в аналитических системах. М.: Флинта, 2019. 207 с.

6. Грановеттер М. Сила слабых связей // Экономическая социология. № 4. 2009. С. 31-50.

7. Губанов Д.А. Социальные сети: модели информационного влияния, управления и противоборства. М.: Изд-во МЦНМО, 2018. 223 с.

8. Данина М.М., Шаляпин А.А. Социально-психологический аспект исследования социальных сетей в Интернете // Вестник Московского университета. Серия 10: Журналистика. № 3. 2012. С. 16-32.

9. Дженкинс, Генри. Конвергентная культура. Столкновение старых и новых медиа пер. с англ. М.: Рипол-Классик, 2019. 384 с.

10. Кастельс, Мануэль. Власть коммуникации / пер. с англ. М.: Изд-во ВШЭ, 2020. 592 с.

11. Методология современных семантических исследований. Коллективная монография. М.: Флинта, 2018. 304 с.

12. Малютин Е. А., Бугайченко Д. Ю., Мишенин А. Н. Выделение текстовых трендов в социальной сети ОК // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2017. Т. 13. Вып. 3. С. 313–325.

13. Социальные изменения в условиях цифровой среды / под ред. В.П Васильева. М.: Макс пресс, 2020. 79 с.

14. Ahmed A., Xing E. P. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream // Proceedings of the Twenty-Sixth Conference. Conference on Uncertainty in Artificial Intelligence. 2010. Vol. 20. P. 29.

15. Bilby, Kenneth. The General: David Sarnoff and the Rise of the Communications Industry. N.Y.: Harper and Row, 1986. 114 p.

16. Burt R. Structural Holes and Good Ideas // American Journal of Sociology. 2019. Vol. 110. P. 349-399.

17. Cross R., Parker A. The Hidden Power of Social Networks. Boston: Harvard Business School Press, 2019. 112 p.

18. Cvijikj I. P., Michahelles F. Monitoring Trends on Facebook // Dependable, Autonomic and Secure Computing (DASC), 2011. P. 895-902.

19. Droogenbroeck F.van. An Essential Rephrasing of the Zipf Law to Solve Authorship Attribution Applications by Gaussian Statistics. URL: https://www.academia.edu/40029629.

20. Hogg, Scott. Understanding and Obey the Laws of Networking: Ignorance of the Laws of Networking Is No Excuse // Network World, October 5, 2013.

21. Peterson, Timothy. Metcalfe's Law as a Model for Bitcoin's Value // Alternative Investment Analyst review, No. 7(2), 2018. P. 9-18.

22. Reed, David P. That Sneaky Exponential: Beyond Metcalfe's Law to the Power of Community Building // Context Magazine. URL: http://www.reed.com/Papers/GFN/reedslaw.html.

23. Schubert E., Weiler M., Kriegel H.-P. Signitrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds // Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014. P. 871-880.

24. Vromen, Ariadne. Digital Citizenship and Political Engagement. URL: https://www.palgrave.com/gp/book/9781137488640.

TECHNOLOGIES OF SOCIAL NETWORKS ANALYSIS TO IDENTIFY SOCIAL TRENDS

Galaganova S.G., Turusina T.V.

Bauman Moscow State Technical University

Abstract. One of the conditions for the effective management of modern society and its subsystems is the timely and accurate determination of social trends - stable trends and dominant directions of social dynamics. In the context of the digital transformation of society, social networks have become the main platform for social communications and, consequently, the collection of empirical data. However, the volume of information in social networks far exceeds the possibilities of its processing in the usual ways, creating a need for algorithmization and automation of this process. The authors of the article made an attempt to develop an analytical algorithm for the subsequent creation of an appropriate software product. The article considers the laws of value of a social network by D. Sarnoff, R. Metcalfe, D. Reed and J. Zipf, analyzes the properties of social networks presented in the form of social graphs, defines metrics that describe the strength of the influence of socio-psychological effects on the user of the network, the specifics of media texts in social networks are revealed, correlations between social effects and the formation of social media trends are derived. The results of the work can serve as a basis for developing a software product for automated processing of a corpus of media texts in social networks.

Keywords. Social network, social trend, trendwatching, semantic analysis, social network value law, social graph, strong ties, weak ties, corpus of texts.