УДК 303.732 **DOI: 10.14451/1.196.160**

МЕТОДОЛОГИЯ СИСТЕМНОГО АНАЛИЗА ИНФОРМАЦИОННОЙ СРЕДЫ

© 2021 **Родионов Дмитрий Григорьевич**
доктор экономических наук, профессор, Высшая инженерно-экономическая школа
Санкт-Петербургский политехнический университет Петра Великого, Россия, Санкт-Петербург
E-mail: dmitry.rodionov@spbstu.ru

© 2021 **Конников Евгений Александрович**
кандидат экономических наук, доцент, Высшая инженерно-экономическая школа
Санкт-Петербургский политехнический университет Петра Великого, Россия, Санкт-Петербург
E-mail: konnikov.evgeniy@gmail.com

© 2021 **Конникова Ольга Анатольевна**
кандидат экономических наук, доцент, кафедра маркетинга
Санкт-Петербургский государственный экономический университет, Россия, Санкт-Петербург
E-mail: olga.a.konnikova@gmail.com

Общемировой тренд цифровизации процессов социального взаимодействия стал первопричиной глобальной многоуровневой и комплексной трансформации как информационной среды в целом, так и паттернов потребления и генерации информации. Данная трансформация привела к появлению единого цифрового информационного пространства — «интернет» — единовременно объединяющим подавляющее большинство потенциальных потребителей, неизменно потребляющих и генерирующих информацию. Следствием данного факта является необходимость формирования новой методологии системного анализа, позволяющей как идентифицировать и агрегировать первичные массивы цифровой информации, так и позволяющие ее квантифицировать и проанализировать. Данная методология может быть потенциально полностью автоматизирована, так как единое цифровое пространство позволяет абстрагироваться от традиционных методов извлечения информации. Однако, общая динамики и бессистемность генерируемой информации неизменного требует привлечения значительно больших аналитических мощностей. Таким образом, целью данного исследования является разработки и методическая детализация методологии системного анализа информационной среды. В качестве предмета исследования была определена новостная информация, как наиболее универсальная составляющая информационной среды.

Ключевые слова: информационная среда, новостная информация, новостная единица, информационный форм, тематический кластер, содержательная компонента, тональная компонента.

В условиях перехода к шестому технологическому укладу процесс цифровизации стал всепроникающим. Одним из наиболее трансформирующихся под воздействием процесса цифровизации элементов социально-экономических систем является информационная среда, под которой следует понимать совокупность информационных условий существования субъекта (человека, предприятия и т.д.). К свойствам трансформации информационной среды социально-экономических систем стоит в первую очередь относить:

1. Значительный рост числа субъектов, обладающих доступом к высокоскоростному интернету. Данное свойство в первую очередь стало следствием повышения доступности высокоскоростного интернета.

2. Экспоненциальный рост скорости распространения информации. Данное свойство является следствием развития технологий связи, в частности интернет-технологий. Повсеместное распространение высокоскоростного интернета определило прирост инвестиций в развитие телекоммуникационных технологий, следствием чего стал значительный прирост скорости передачи цифровой информации.

3. Значительный рост объемов передаваемой цифровой информации за единицу времени. Вследствие повышения доступности высокоскоростного интернета число пользователей значительно динамично увеличивается с каждым годом, что определяет прирост объемов пе-

редаваемой информации.

4. Переход значимой части маркетинговой и торговой активности в цифровую среду. Следствием повышения скорости и доступности высокоскоростного интернета стало повышение среднего времени, проводимого пользователями «в сети», что, в сочетании с общим приростом пользователей, определило необходимость интеграции бизнеса (в первую очередь торгового бизнеса) в единое цифровое информационное пространство, с целью анализа, формирования и управления потребностями потребителей. Результатом данной интеграции стало появление уникальных направлений деятельности — электронный маркетинг и электронная коммерция. В соответствии с исследованием компании Data Insight в 2020 году объем российского рынка электронной коммерции составил 2,5 триллиона рублей, что на 44% больше в сравнении с 2019 годом. Безусловно, данная динамика во многом обусловлена пандемией COVID-19, однако, исследовательская группа отмечает, что без учета влияния пандемии COVID-19, рост должен был составить 29% [1].

5. Развитие уникальных коммуникативных паттернов в рамках цифровой информационной среды. Вследствие повышения среднего времени, проводимого интернет-пользователями «в сети», в сочетании с общим приростом интернет-пользователей, значительный объем социальной коммуникации также был интегрирован в единое цифровое информационное пространство, следствием чего стало развитие особых цифровых концентраторов интернет-пользователей — социальных сетей. В наиболее широком смысле под социальной сетью следует понимать любое сообщество, организованное едином цифровой информационном пространстве, участники которого имеют возможность в различных формах реализовывать процесс коммуникации. В качестве наиболее общих функций социальных сетей следует выделять возможность создания индивидуальных цифровых профилей, содержащих какую-либо информацию о пользователе и возможность взаимодействия пользователей. По данным на 2018 год 78% населения жителей России состояло хотя бы в одной социальной сети, что являлось вторым результатом в мире (первое место занимала Япония, с показателем в 88%) [2]. Обозначенные уникальные коммуникативные паттерны в первую очередь проявляются как установлении цифровой коммуникации как первичной, а также в формировании и дифференциации множественных цифровых профилей, определяющих характер коммуникации и самопрезентации в зависимости от выбранного цифрового коммуникативного пространства.

6. Формирование уникальных осознанных цифровых профилей интернет-пользователей. Как выделялось ранее, коммуникация интернет-пользователей в социальных сетях как правило подразумевает формирование цифровых профилей, выступающих в роле презентационных инструментов, и отражающих сформированный пользователем образ. Данный цифровой профиль может значительно дифференцироваться у одного пользователя, в зависимости от выбранной социальной сети, однако, так как значительная часть элементов данного профиля (в первую очередь комментарии в тому или иному цифровому контенту) формируется под воздействием моментного эмоционального фона, пользователь как правило не в состоянии обеспечить перманентную рефлексию генерируемого материала, что в свою очередь позволяет формулировать достоверные выводы на основе анализа данной информации.

7. Формирование уникального «цифрового следа» интернет-пользователя. Под цифровым следом стоит понимать генерируемые пользователем в сети интернет метаданные, связанные со временными и пространственными характеристиками перемещения пользователя в едином цифровом пространстве за анализируемый период. Данный цифровой след как правило является неосознанным и на основе его аналитической обработки маркетологи в частности получают возможность настраивать свойства контекстной рекламы индивидуально для потенциального потребителя или группы потенциальных потребителей.

8. Децентрализация субъектов управления процессом генерации и распространения информации. Следствием всего вышесказанного стал переход от системы полностью или частично управляемой к полностью децентрализованной системе генерации и распространения информации. На данный момент каждый из интернет-пользователей имеет возможность генерации общедоступного контента, который является частью единой цифровой информационной среды, доступным к распространению и дискуссии, в рамках которой также может гене-

рироваться общедоступный контент.

Выделенные свойства определяют актуальность использования генерируемой в единой цифровой информационной среде информации в качестве аналитического базиса, для целей формирования и уточнения управленческих решений, при одновременной ограниченности управленческого воздействия на данную информационную среду и ее субъектов со стороны государственных органов и представителей бизнеса. На данный момент именно в рамках цифровой информационной среде непрерывно генерируется аналитическая информация, наиболее точно и в полном объёме отражающая свойства потенциальных потребителей той или иной продукции, определяющая образ восприятия того или иного предприятия, позволяющая установить сравнительную актуальность той или иной тематики и т.д. При этом цифровая информационная среда трансформируется непрерывно, и под воздействием как внешней среды, так и под воздействием внутренней, отражая не только свойства и процессы, происходящие в материальном мире, но и собственные свойства и процессы, а также процессы метаосмысления трансформации. Таким образом, описание и анализ цифровой информационной среды требует системного подхода. В рамках данного исследования рассматриваются исключительно локальные аспекты системного анализа информационной среды, ограниченные анализом общего новостного фона. Однако, необходимо отметить, что принципы, заложенные в разрабатываемые инструменты, могут быть экстраполированы на решение смежных задач анализа как информационной среды в целом, так и ее частных объектов.

Концепция воздействия новостного фона, как элемента информационной среды, на человека определяется спецификой массового восприятия. Как отмечалось ранее, информация, размещенная в единой цифровой информационной среде, может быть доступной каждому интернет-пользователю, что определяет условную массовость восприятия, и как следствие воздействия данной информации на представителей социума. Массовость является условной, так как решение относительно потребления той или иной информации принимает сам пользователь, в первую очередь определяя посещаемые им информационные ресурсы. Однако, новостная информация является одной из наиболее проникающих. Ингрессия новостной информации в цифровой информационной среде определяется в первую очередь их актуальностью и оперативностью, что в свою очередь значительно увеличивает потенциальный охват аудитории и ее выход за пределы стандартных тематических блоков. Таким образом, именно новостная информация может оказывать потенциально наиболее разнонаправленное, с точки зрения аудитории, воздействие на социум. Отражая объективные изменения во внешней для потребителя информации среде, специфика новостного фона может относительно эффективно воздействовать как на его эмоциональное состояние. Эмоции, являясь отражением субъективно-оценочного отношение субъекта к тем или иным событиям объективного мира, во многом определяет вектор воздействия субъекта на объективный мир. При этом, так как эмоции являются психическим процессом исключительно короткой или средней продолжительности, последствия данного воздействия во многом носят локальный характер [3]. Одним из наиболее значимых (с точки зрения частоты проявления) типов воздействия на объективный мир в данном случае является генерация ответного информационного потока в рамках цифровой информационной среды. Данный поток может мультиплицироваться вследствие распространения первичной новостной информации, как компоненты данного информационного потока, что в свою очередь порождает экспоненциальный рост потребителей данной новостной информации. Таким образом, потенциальная массовость распространения той или иной новостной информации в рамках цифровой информационной среды во многом определяется силой ответной реакции аудитории на данную информацию. Также, помимо эмоциональной составляющей необходимо отметить устойчивую содержательную составляющую. Именно окружающий информационный фон во много переделяет восприятие человеком объективного мира, а новостной фон является значимой частью информационного фона. Таким образом, системный анализ новостной информации позволит прогнозировать как кратко и среднесрочные эмоциональные трансформации исследуемой аудитории, так и формировать стратегии управления восприятием объективного мира.

Новостная информация в первую очередь представлена в закодированной форме, пред-

ставляющей собой текст, изложенный естественным языком. Безусловно, данный текст может сопровождаться потоком аудиовизуальной информации, однако она является вторичной по отношению к текстовой. Данная специфика характерна в первую очередь для новостной информации, так как она как правило содержит множественно последовательных тезисов, раскрывающих в первую очередь формальную составляющую актуального события. Информация, представленная посредствам естественно языка, крайне специфичная с точки зрения анализа. С точки зрения концептуальных компонент, обладающих аналитической ценностью, могут быть выделены содержательная и тональная компоненты [4]. Тональная компонента отражает эмоциональный окрас представленной новостной информации. Типов эмоций в психологии крайне много, как и классификационных признаков их разделяющих, однако, в рамках инструментального анализа естественного языка традиционно выделяют минимум три эмоциональные характеристики текстовой информации — нейтральность, негативность и позитивность. Безусловно, в рамках отдельных инструментов выделяются дополнительные характеристики, однако именно данные три характеристики составляют универсальный базис оценки тональной компоненты естественной текстовой информации. Значимость исследования тональной компоненты определяется в первую очередь потенциальным воздействием новостной информации на человека посредствам механизмов эмпатии. Переходя на более обобщённый аналитический уровень, можно утверждать, что при достаточной распространённости тонально смоделированного новостного фона, субъект управления получает возможность воздействовать на общие эмоциональные характеристики массовой аудитории, что в свою очередь может оказывать влияние на социальные, политические и экономические показатели социально-экономических систем. Содержательная компоненты отражает наличие в новостной информации той или иной тематической составляющей. Данная компонента может дифференцироваться в зависимости от потребностей исследователя. Управление наличием той или иной содержательной компоненты в новостной информации дает возможность субъекту управления определять и в долгосрочной периоде моделировать вектор развития тематических интересов социума.

Методология анализа содержательной и тональной компоненты является смежной, и в основе нее лежит принцип токенизации исследуемой текстовой информации. С точки зрения прикладной реализации, первичным этапом данного процесса выступает идентификация источников новостной информации, соответствующим следующим характеристикам:

1. Достаточный охват аудитории. Достаточность я данном случае является нечетким параметром, и может варьироваться в зависимости от специфики решаемой задачи. Однако, определить достаточном возможна посредствам сравнительной оценки двух параметров: дифференциации аудитории и широты аудитории.

2. Объективная дифференциация и отсутствие единого вектора изложения. Данная характеристика подразумевает отсутствие универсализированного оценочного взгляда авторов и учредителей на трактовку новостной информации. Максимизация данного параметра подразумевает максимизацию объективности излагаемой информации.

3. Хронологическая структурированность. Представленная новостная информация должна быть представлена в виде временного ряда достаточной длинны, универсализированная с точки зрения формы изложения. Максимизация длинны временного ряда позволит строить прогностические модели для квантификационных параметров новостной информации.

4. Информационная актуальность. Представленная новостная информация должна быть наиболее актуальной, и фактически отражать текущее состояние внешней среды подавляющего большинства интернет-пользователей.

5. Пригодность для автоматизированной обработки. Данная характеристика является дополнительной и формируется на основе методологической специфики. Вне наличия возможности автоматизированного сбора и обработки новостной информации исследователь вынужден затрать в рамках данного этапа время, несовместимое с процессом обновления информационного фона, что в свою очередь не позволит актуализировать результаты анализа.

Представленная совокупность характеристик определяет необходимость первичного экспертного отбора источников новостной информации. По результатам сравнительной оценки и определения совокупности данных источников производится формирование уникального ин-

струмента автоматизированного сбора и структурирования новостной информации. Данный этап может быть реализован средствами инструментов, разработанных на языке программирования Python 3. Данный выбор в первую очередь обусловлен значительным объёмом универсальных решений в сфере как парсинга, так и обработки цифровых данных. К инструментам автоматизированного сбора цифровых данных можно отнести библиотека selenium и requests, позволяющие отправлять http запросы и расшифровывать полученный ответ. Полученная информация представлена в форме html кода советующего интернет-ресурса, из которого необходимо извлечь хронологическую и текстовую информацию, описывающую новостной фон.

По результатам формирования описанных временных рядов, производится первичная обработка, трансформирующая естественного текста в массивы приставляющих его лексем — токенов. Однако, сформированные первичные массивы токенов непригодны для анализа, так как они не универсализированы со словарной и регистровой точки зрения, а также содержат множества информационно-ненасыщенных лексем, таких как знаки пунктуации, союзы, предлоги, местоимения, устойчивые выражения и т.д. Таким образом последующая обработка массивов токенов может быть дифференцирована на два базовых последовательных этапа:

1. Лемматизация — процесс приведения токена к его словарной форме [5]. По результатам данного этапа формируется значительно более оптимизированный массив токенов, содержащий исключительно их словарные формы в нижнем регистре.

2. Исключение информационно-ненасыщенных лексем. Данный этап подразумевает предварительное формирование словаря, содержащего данные лексемы. По результатам исключения данных лексем массив токенов также значительно поэтизируется.

Описанный алгоритм графически представлен на рисунке 1. В рамках данного исследования апробация представленных алгоритмов производится на основе материалов, представленных генераторами новостной информации в социальной сети «ВКонтакте» (vk.com). Данный выбор обусловлен в первую очередь наличием функционального и доступного API интерфейса, позволяющего агрегировать всю официально представленную и открытую для пользователей информацию, размещенную как в социальных профилях пользователей социальной сети, так и в официальных цифровых сообщества пользователей. Значительная популярность данной социальной сети определяет необходимость организации новостными генераторами официальных сообществ в рамках данной социальной сети, полностью или частично дублирующих

Рисунок 1. Алгоритм идентификации, сбора и токенизации новостной информации

новостную информацию. В рамках данного исследования, в качестве источника новостной информации было определено официальное сообщества информационного агентства «Вести». Данный выбор обусловлен тем, что данное информационное агентство в полном объеме и своевременно дублирует новостную информацию, представленную на их официальном сайте, что позволяет утверждать о потенциальном соответствии данного источника все заявленным ранее характеристикам. На рисунке 2 представлен детализированный алгоритм сбора новостной информации в официальном сообществе информационного агентства «Вести» в социальной сети «ВКонтакте».

По результатам применения данного алгоритма для извлечения 25% представленной новостной информации, 32 601 новостная единица, насыщенная как естественной текстовой информацией, так и метаинформацией. Далее необходимо систематизировать извлеченную информацию и агрегировать исключительно необходимый свод элементов. Архитектура API исследуемого информационного ресурса позволяет извлечь следующие значимые данные:

1. Текст новостной единицы. Данная информация является наиболее значимой, и фактически содержит, представленные в естественной форме, исследуемые информационные компоненты.

2. Дата размещения новостной единицы в открытом доступе. Данная метаинформация выступает ключевым классификационным параметром, позволяющим сформировать на основе сформированным массивов временные ряды, а также в дальнейшем усреднить значения исследуемых параметров в рамках заданных временных периодов.

3. Число «комментариев» пользователей для каждой новостной единицы. Сущностно, комментарии являются индикатором реакции пользователей на содержание новостной единицы, что позволяет проводить сравнение новостных единиц относительно формирования общественного отклика/резонанса.

4. Число «лайков» пользователей для каждой новостной единицы. Данный параметр отражает степень условного одобрения или согласия пользователей с содержанием новостной единицы. Данный параметр во многом может быть использован для определения наиболее коррелирующих с восприятием аудитории тематических компонент, тональных характеристик или, на метауровне, форм отражения актуальной информационной повестки.

5. Число «репостов» пользователей для каждой новостной единицы. Данный параметр является развитием параметра «лайк», и характеризует желание пользователей ретранслировать отраженное в рамках новостной единицы сочетание содержательных и тональных свойств, которые также неизменного должны рассматриваться в контексте актуальной информационной повестки.

6. Число «просмотров» пользователей для каждой новостной единицы. Данный параметр является наиболее общим, и он отражает уровень распространённости новостной единицы в рамках существующей аудитории.

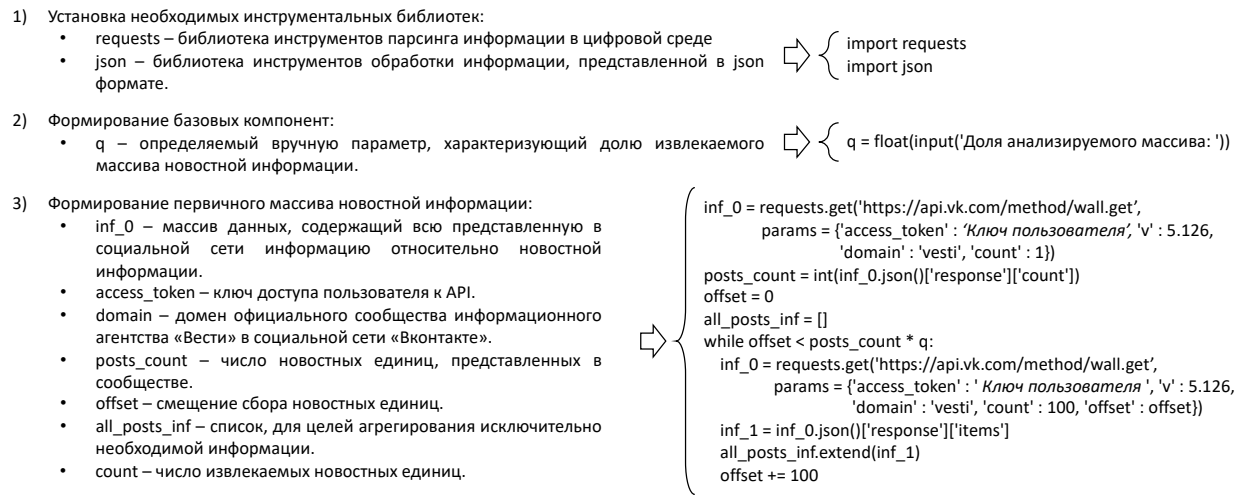

Рисунок 2. Алгоритм сбора новостной информации в официальном сообществе информационного агентства «Вести» в социальной сети «ВКонтакте»

Реализация алгоритма структурирования извлеченной новостной информации, на языке программирования Python, представлена на рисунке 3.

По результатам реализации данного алгоритма формируется 6 сопоставимых одномерных массивов аналитической информации. Также, необходимо отметить возможность формирования вторичной аналитической информации, являющейся производной от представленной информации. В частности, могут быть рассчитаны следующие аналитические показатели:

1. Отношение число лайков к числу просмотров новостной единицы. Данный параметр отражает первый уровень эмоциональной конверсии восприятия содержания новостной единицы. Относительно большее значение данного показателя указывает на сравнительной большую корреляцию представления анализируемой новостной единицы с восприятием содержания аудиторией пользователей. Данный показатель определяется маркетологами как «показатель вовлеченности аудитории».

2. Отношение числа репостов к числу просмотров новостной единицы. Данный показатель является развитием предыдущего показателя и также отражает эмоциональную конверсию восприятия содержания новостной единицы. Относительное превышение данного показателя отражает степень концентрации вовлеченной аудитории.

3. Отношение числа комментариев к числу просмотров новостной единицы. Данный показатель отражает относительный эмоциональный отклик аудитории на содержание новостной единицы. При этом, данный эмоциональный отклик может носитель как положительный, так и отрицательный характер. Возможный дуализм восприятия содержания новостной единицы может значительно прирастить данный показатель, за счет формирования дискуссии между пользователями.

Анализ распределения значений представленных производных показателей позволит идентифицировать совокупности новостных единиц, вызывающих наибольшую вовлеченность аудитории и наибольший эмоциональный отклик. В рамках завершающего подэтапа первичного этапа методики автоматизированного системного анализа новостной информации производится токеннизация текстов извлеченных новостных единиц. Как отмечалось ранее, данных этап подразумевает последовательную обработку текстовой информации. Алгоритм токенизации извлеченной новостной информации представлен на рисунке 4.

Полученных массивы токенов, распределенные в соответствии с их принадлежностью к хронологически структурированных новостным

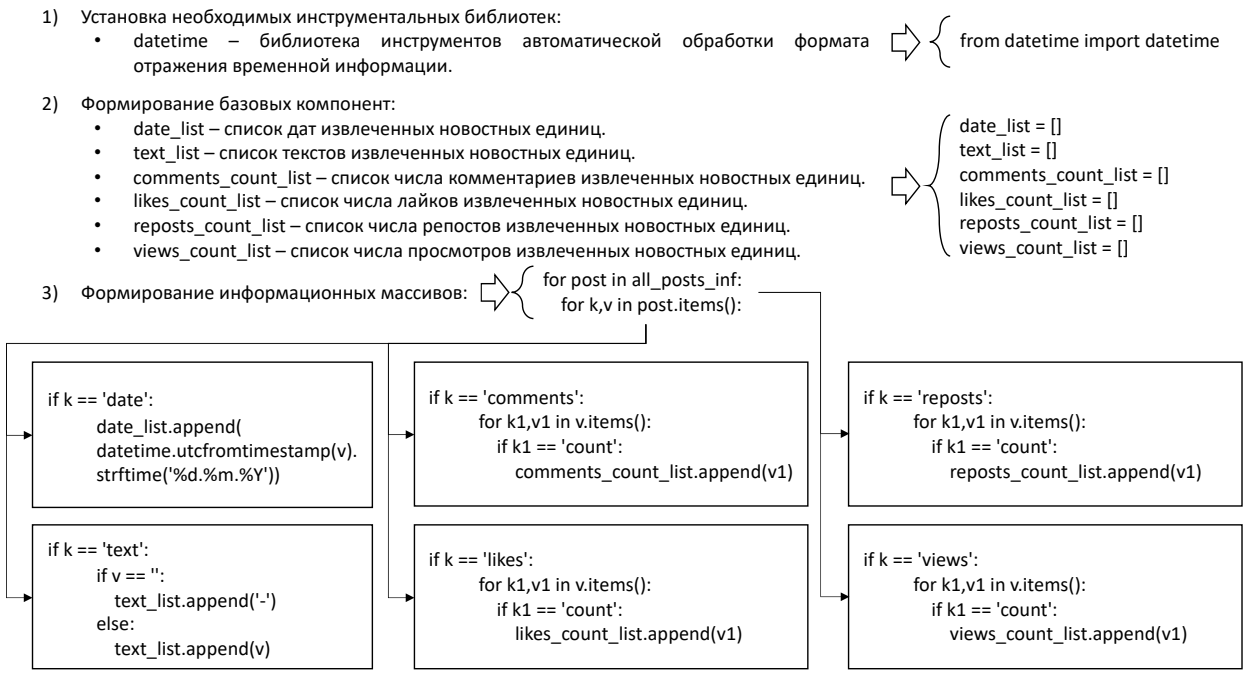

Рисунок 3. Алгоритм структурирования извлеченной новостной информации

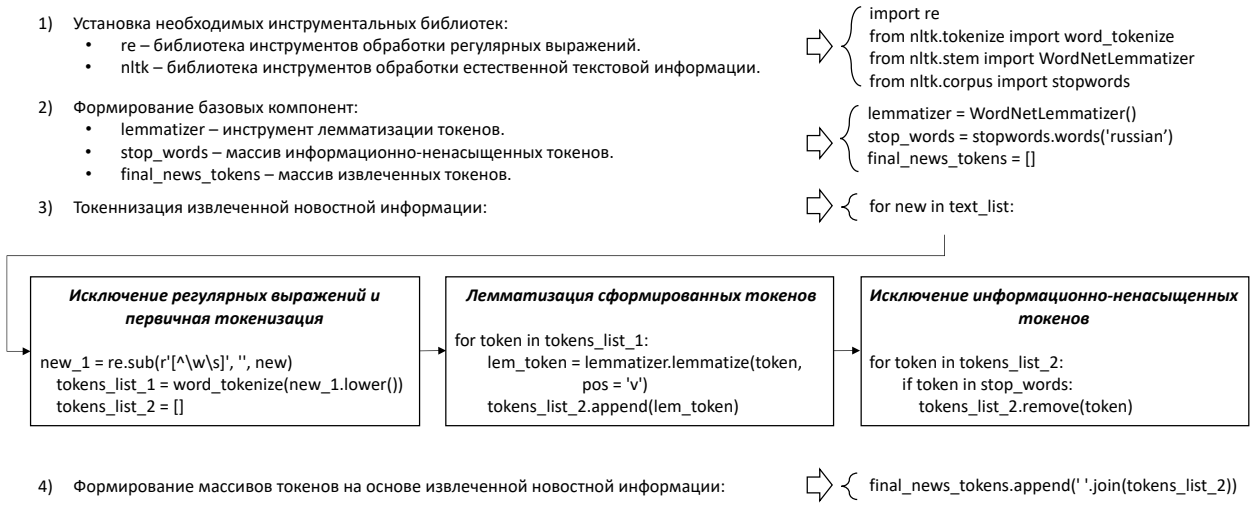

Рисунок 4. Алгоритм токенизации извлеченной новостной информации

единицам являются основой для анализа содержательной и тональной компонент извлеченной новостного информации. Рассмотрим в первую очередь анализ содержательной компоненты. Данная компонента является крайне дифференцированной и не подразумевает универсализации вне контекста выделения конкретных параметров содержания. Выделение конкретных параметров содержания определяется исследовательской необходимостью, и представляет собой дискретно установленных массив токенов, соответствующих той или иной тематики. В частности, при необходимости выделения новостей по конкретной, дискретно заданной тематики, необходимо организация качественного исследования, подразумевающего в первую очередь привлечения соответствующих экспертов, результатом взаимодействия с которыми является соответствующих массив тематических токенов. Пример автоматизированной методики оценки присутствия дискретно определенной содержательной компоненты в структурированном массиве новостной информации детально представлен в работе [6]. В рамках данного исследования рассматривается значительно более функциональный и универсальный алгоритм, подразумевающий первичное автоматизированное формирование тематических кластеров, подразумевающее последующую оценку присутствия каждого из выделенных кластеров в каждой из извлеченных новостных единиц, или их совокупностей. В качестве описательной основы предлагается использовать математическую модель «Bag of words» или «Мешок слов». Данная модель была разработана в 1975 году Дж. Солтоном и представляет собой матрицу, каждый из столбцов которой определяется одним из выделенных ранее токенов, а каждая из строк — текстовой единицей, потенциально содержащей данные токены, и использованной в рамках их извлечения. В рамках данного исследования в качестве текстовых единиц выступают извлеченные новостные единицы. На пересечении токенов и новостных единиц размещается частота встречаемости конкретного токена в конкретной новостной единице. В рамках данной модели методологически не может быть выделена эндогенная переменная, однако, факт принадлежности токена к конкретной новости и частота его встречаемости позволяет реализовать кластерный анализ, и выделить фиксированное число кластеров, объединяющих наиболее часто совместно встречающихся токенов. Сформированные массивы квантифицируются уровнем принадлежности каждого из токенов конкретному кластеру. При этом, необходимо отметить, что каждый из токенов может быть сравнительно единообразно быть значим для нескольких кластеров. Каждый из выделенных кластеров подлежит экспертной обработке, подразумевающей его селекцию, выделение уникальных, наиболее информационно-насыщенных токенов и детализированное тематическое описание каждого из скорректированных кластеров. Каждый из результирующих кластеров представляет собой упомянутую ранее уникальную дискретно определенную содержательную компоненту. В рамках завершающего этапа алгоритма анализа содержательной компоненты новостной информации предполагается определение доли

наличия содержания каждого из выделенных тематических кластеров в каждой из извлеченных новостных единиц, что может быть реализовано посредствам последовательного соотнесения двух множеств. На рисунке 5 представлен алгоритм анализа содержательной компоненты новостной информации. Рассмотрим пример реализации данного алгоритма более детально, применительно к сформированному ранее массиву токенов. В первую очередь необходимо детально рассмотреть автоматизацию этапа определения числа тематических кластеров. Значительно часть инструментов, используемых в рамках данного этапа, автоматизирована в рамках Python библиотеки sklearn. Процесс формирования модели «Bag of words» на основе сформированного массива токенов, автоматизировано в рамках инструмента CountVectorizer. Полученная частотная матрица далее циклически обрабатывается выбранным инструментом кластеризации, для каждого из возможным значений числа кластера. Для целей кластеризации сформированного массива новостной информации, использован метод k-средних, основанный на принципе минимизации суммы квадратов отклонений точек кластеров от центральных позиций данных кластеров. Выбор метода k-средних в первую очередь обусловлен сравнительно высокой скоростью обработки информации. По результатам кластеризации рассчитывается силуэтная оценка, отражающая степень подобия точки данных с собственным кластером по сравнению с иными сформированными кластерами, и определяющая качество кластеризации. По результатам идентификации наибольшего значения силуэтной оценки, определяется соответствующее данной оценке число кластеров. Детализированный алгоритм тематической кластеризации извлеченной новостной информации представлен на рисунке 6.

Для исследуемого массива новостной информации наибольшее значение силуэтной оценки было достигнуто при выделении одиннадцати тематических кластеров. Каждый из выделенных кластеров требует экспертной обработки, представляющей собой детальное осмысление содержания. Рассмотрим последовательно каждый из выделенных в рамках описываемого

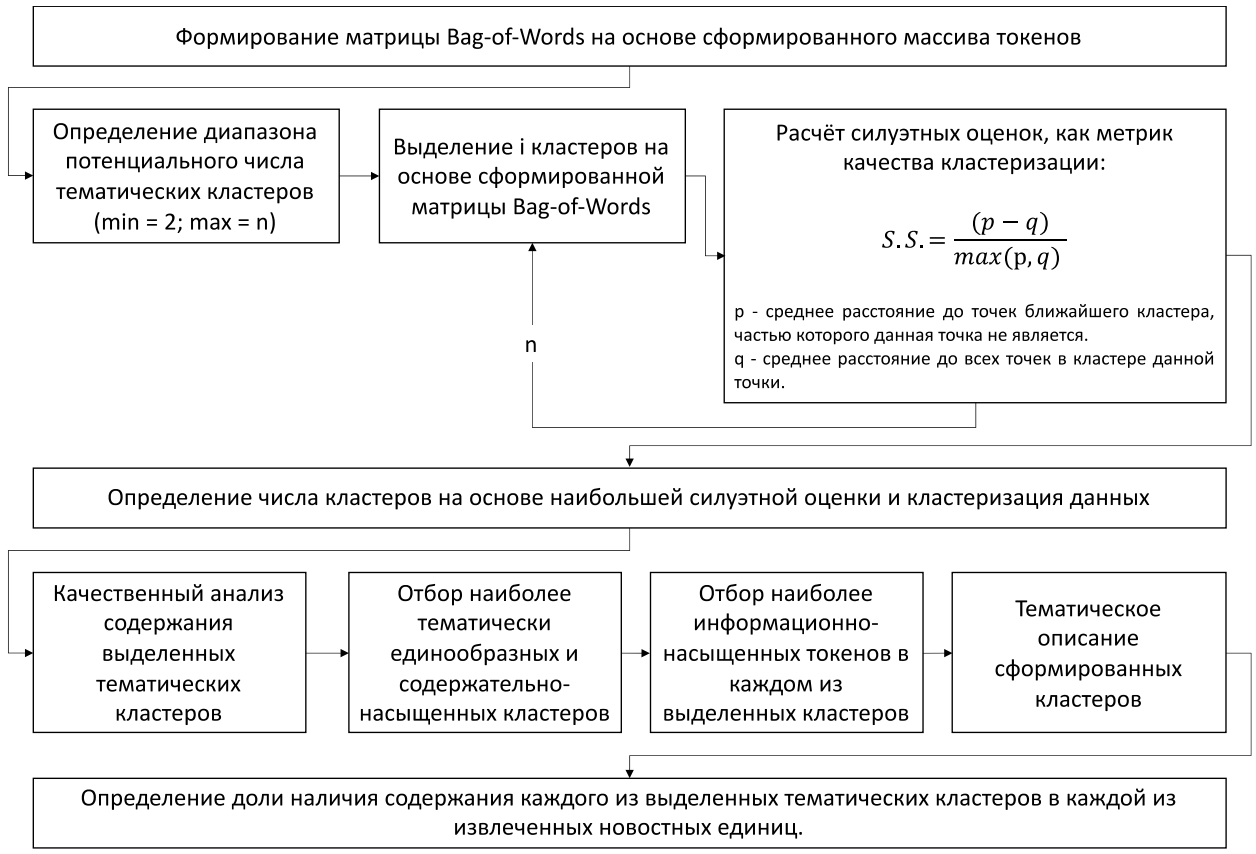

Рисунок 5. Алгоритм анализа содержательной компоненты новостной информации

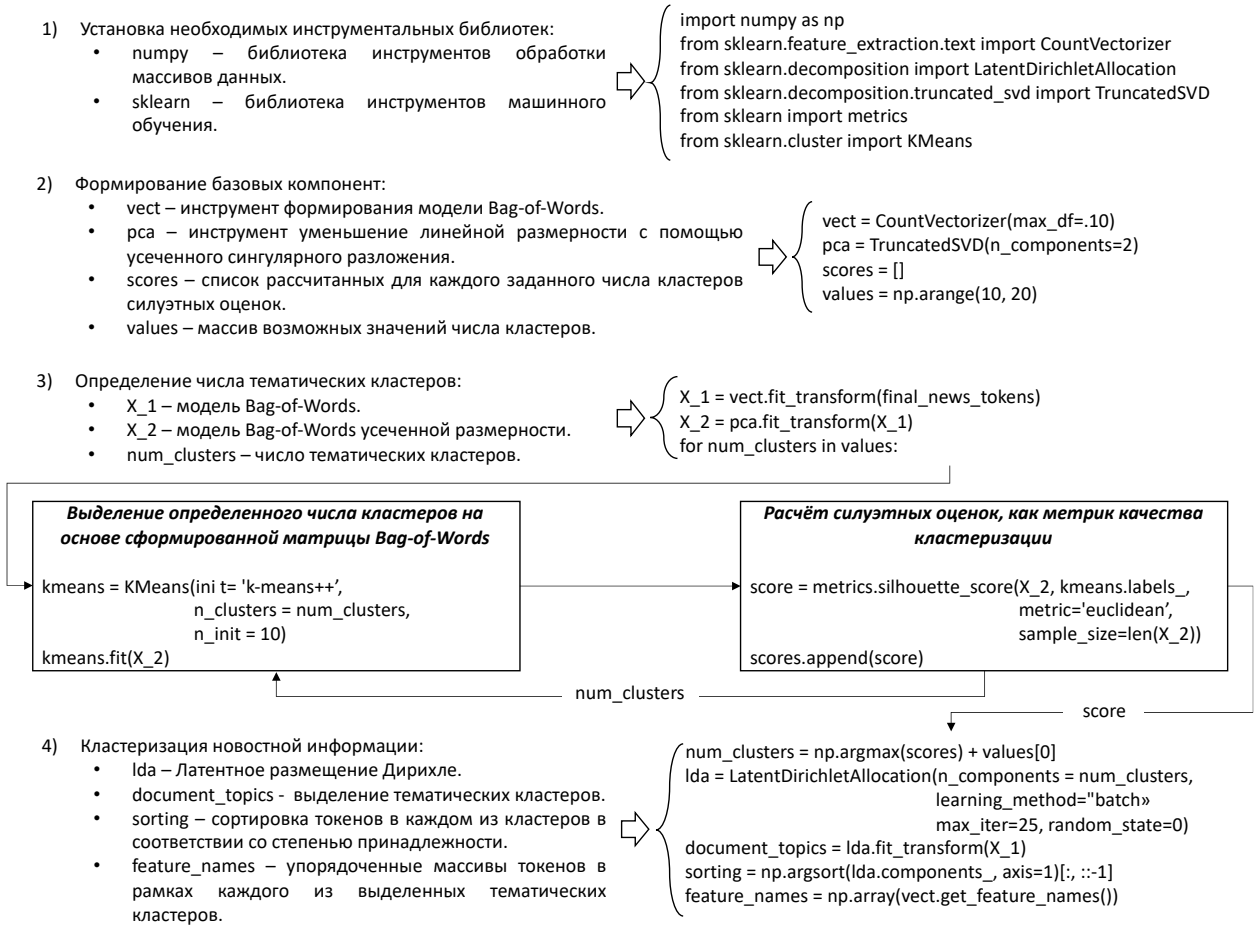

Рисунок 6. Алгоритм тематической кластеризации извлеченной новостной информации

примера одиннадцати кластеров:

1. “Кластер 1” определяется такими токенами как: “США”, “северный”, “сегодня”, “дело”, “поток”, “деньги” и т.д. На основе анализа содержания, можно предположить, что данный кластер объединяет новостные единицы, посвященные международным санкциям и в целом международной реакции (в частности реакции США) на строительство “Северного потока 2”. Так как анализу подвергся массив новостной информации, сформированный исключительно в рамках 2019 и 2020 года, отдельное выделение данного кластера является обоснованным, так как данный международный проект является наиболее масштабным для экономики России, что определяет интерес к международного сообщества к использованию его в качестве некого общественно-политического рычага воздействия. Несмотря на относительную локальность, данный кластер может быть использован в рамках дальнейшего анализа.

2. “Кластер 2” определяется такими токенами как: “рублей”, “суд”, “тысяч”, “лет”, “миллионов”, “деньги”, “долларов”, “миллиона”, “задержали”, “дело и т.д. На основе анализа содержания, можно предположить, что данный кластер объединяет новостные единицы, посвященные экономическим преступлениям и правонарушениям. В данном случае, можно предположить, что отдельное выделение данного кластера обусловлено значительным количеством новостей о коррупции и процесса противодействия ей. Может быть выдвинута гипотеза, относительно потенциально более значимой общественной реакции на новостные единицы, соответствующие данному кластеру, что определяется необходимость его использования в рамках дальнейшего анализа.

3. “Кластер 3” определяется такими токенами как: “coronavirus”, “covid19”, “коронавирус”, “число”, “случаев”, “день”, “тысяч”, “мире”, “заболевших” и т.д. Данная совокупность токенов в полной мере определяет тематику процесса и последствий распространения вируса COVID-19.

Данная тематика является чрезвычайно актуальной, ее превалирование в информационном пространстве в значительной мере характеризована 2020 год. Соответствующая специфика позволяет выдвинуть термин "Информационное распространение коронавирус", определяющий степень охвата информационной среды соответствующей тематикой. Безусловно, отдельное выделение данного тематического кластера целесообразно в рамках дальнейшего анализа.

4. "Кластер 4" определяется такими токенами как: "Карабах", "Армения", "Азербайджан", "Нагорный" и т.д. Данный кластер объединяет в себе новостные единицы, посвященные конфликту между Арменией и Азербайджаном в Нагорном Карабахе во второй половине 2020 года. Также, надо отметить, что данный кластер насыщен множеством универсальных относительно информационно-ненасыщенных токенов, что позволяет судить о нем, как о условно незначимом. Более того, хронологическое расширение новостного массива вероятнее всего исключит данный кластер. Таким образом, использование данного кластера в рамках дальнейшего анализа нецелесообразно.

5. "Кластер 5" определяется такими токенами как: "coronavirus", "россиян", "рублей", "тысяч", "коронавируса" и т.д. В рамках данного кластера объединяется экономическая тематика и тематика, основанная на пандемии коронавируса. Следовательно, данный кластер можно условно охарактеризовать как "экономические последствия пандемии коронавируса". Так как в дальнейшем предполагается калькуляция доли присутствия каждой из тематик в каждой из новостных единиц, отдельное выделение данной комплексной тематики нецелесообразно, определяет отсутствие необходимости использования данного кластера в рамках дальнейшего анализа.

6. "Кластер 6" определяется такими токенами как: "человек", "области", "жители", "результате", "погибли", "пострадавших", "районе" и т.д. В соответствии с выделенными токенами, может быть выдвинуто предположение, что в рамках данного кластера выделяются новости, посвященные локальным авариям и техногенным или климатическим бедствиям. Данная тематика является относительно перманентной с точки зрения хронологического проявления, однако, возможные последствия могут масштабироваться за счет нарушения устойчивых экономических, логистических и социальных связей, в связи с чем выделение данного кластера в рамках дальнейшего анализа целесообразно.

7. "Кластер 7" определяется такими токенами как: "смотрите", "фильм", "новый, "шоу", "видео", "день", "программы", "кино", "выпуск" и т.д. В рамках данного кластера выделяется новостные единицы, посвященные развлекательному сегменту media-индустрии. Следует выдвинуть гипотезу о том, что данная тематика оказывает незначительное влияние на социальную и экономическую среду, в связи с чем выделение данного кластера в рамках дальнейшего анализа нецелесообразно.

8. "Кластер 8" определяется такими токенами как: "вечер", "Соловьев", "Владимир", "жизни" и т.д. Данный кластер является исключительно локальным, так как объединяет новостные единицы, посвященные и сгенерированные в рамках телевизионные проектов Владимира Соловьева. Выделение данного кластера обусловлено источником новостной информации, а его использование в рамках дальнейшего анализа крайне нецелесообразно. Более того, данный кластер также наполнен информационно-ненасыщенными токенами, таких как "новость", "из-за" и т.д.

9. "Кластер 9" определяется такими токенами как: "мужчина", "женщина", "ребенка", "дома", "мать", "детей", "удалось" и т.д. Данный кластер объединяет новостные единицы, посвященные семейной тематике. Безусловно, соответствующая тематика является исключительно общей и не оказывает значительного воздействия на социальную и экономическую среду, в связи с чем выделение данного кластера в рамках дальнейшего анализа нецелесообразно.

10. "Кластер 10" определяется такими токенами как: "Путин", "президент", "Владимир", "США", "заявил", "РФ", "Россия", "глава", "Украина", "страны", "государства" и т.д. Данный кластер является тематически широким, однако критерии, определяющие принадлежность новостных единиц к данному кластеры, исключительно конкретные. В рамках данного кластеры объединена новостная информация, посвященная деятельности В.В.Путина, в частности в рамках международной политики. Данная тематика определяет глобальные трансформации во внешней среде, что говорит о необходимости выделения данного кластера в рамках дальнейшего анализа.

11. “Кластер 11” определяется такими токенами как: “сутки”, “последние”, “человек”, “данные”, “мира”, “день”, “период”, “прошедшие” и т.д. Данный кластер объединяет в себе новостные единицы, характеризующиеся оперативностью и актуальностью. При этом тематическое единообразие в рамках данного кластера не достигается. При общей широте данного кластера его присутствие в новостной информации может определяться как вторичное свойство, тем самым усиливая или ослабляя свойства иных кластеров. Таким образом, использование данного кластера в рамках дальнейшего анализа целесообразно.

Необходимо отметить, что выделенные тематические кластеры могут значительно трансформироваться при увеличении или уменьшении анализируемого новостного массива. Для целей расширения потенциальных результатов анализа выводы относительно значимости отдельных тематических кластеров, в рамках данного исследования, проигнорированы, и далее при анализе используется каждый из выделенных кластеров. На завершающем этапе описания содержательной компоненты новостной информации производится оценка присутствия каждого из тематических кластеров в извлеченных новостных единицах. Как отмечалось ранее, для целей оценки данного параметра может быть использована частотная интерпретация пересечения двух множеств: множества токенов, образующих новостную единицу и множества токенов, образующих тематический кластер. Таким образом доля того или иного тематического кластера в содержании новостной единицы определяется как отношение числа общих для множества, образующего кластер, и множества, образующего новостную единицу, токенов к сумме токенов, общих для всех кластеров и новостной единицы:

$$D_{t.c._{j_i}} = \frac{\sum\left(t.c._1^m{}_j \cap t.new._i\right)}{\sum_{j=1}^{n} N_{t.c._j}} \quad (1)$$

Где:

1. $t.c._1^m{}_j$ — множество токенов, образующих кластер j (от наиболее значимого до m).
2. m — переделённое число токенов, извлеченных из анализируемого кластера (кластер упорядочен по уменьшению соответствия токенов анализируемому кластеру).
3. $t.new._i$ — множество токенов, образующих новостную единицу i.
4. $N_{t.c._j}$ — число токенов, единых для множества токенов, образующих кластер j и множества токенов, образующих новостную единицу i.
5. $D_{t.c._{j_i}}$ — доля тематического кластера j в новостной единице i.

Необходимо отметить, что определенный ранее параметр соответствия токена тому или иному кластеру, может быть использован в качестве удельного веса, что в свою очередь потенциально позволит уточнить результаты калькуляции. В рамках данного исследования учет удельного веса не производится, так как выделенные кластеры носят условный характер и не подвергаются углубленной вторичной селекции. Отдельно необходимо исследовать параметр m — предельное число токенов, извлеченных из анализируемого кластера. Данный параметр является вариативным, и определяется исключительно исследователем. Так как объём кластеров единой, и они отличаются исключительно иерархически, отказ от использования удельного веса определяет необходимость интеграции и обоснования данного параметра. Приращение данного параметра характеризуется сглаживанием границ между кластерами, тем самым обеспечивая более равномерное распределение удельного веса. При необходимости более повышения содержательной категоричности кластеров, текущий параметр необходимо сокращать, однако, следствием данного сокращения может стать относительное обнуление доли части кластеров, что в свою очередь может отрицательно сказаться на результатах дальнейшего анализа. В рамках данного исследования, в качестве параметра m установлено значение 20. Выбор данного значения обусловлен отсутствием углубленной вторичной селекции токенов, что в свою очередь привело к наличию значительного количества общим для всех кластеров информационно-ненасыщенных токенов. По результатам калькуляции данного параметра формируется матрица размерностью C.N. на N.N., где C.N.— число выделенных кластеров, а N.N.— число извлеченных новостных единиц. Детально, автоматизированный алгоритм присутствия тематических кластеров в извлеченных новостных единицах представлен на рисунке 7. Определенный по результатам последовательной реализации представленного алгоритма параметр является квантификатором содержа-

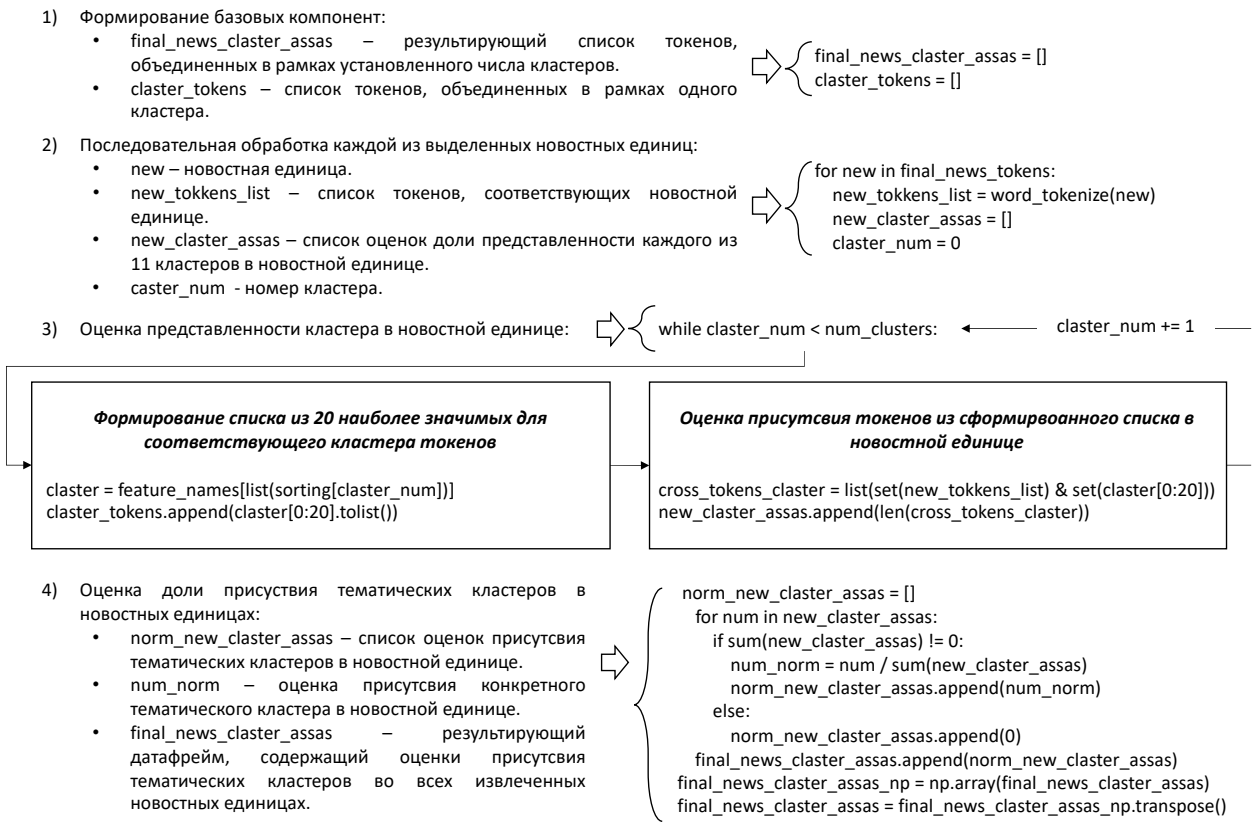

Рисунок 7. Алгоритм оценки присутствия тематических кластеров в извлеченных новостных единицах

тельной компоненты новостной информации, и его многомерный и комплексный анализ позволит установить значимость и специфику влияния каждого из выделенных содержательных кластеров как на параметры внутренне среды новостной информации, так и на параметры внешней среды. Далее необходимо рассмотреть алгоритм оценки тональной компоненты извлеченных новостных единиц. С сущностной и методической точек зрения процесс оценки данной компоненты является идентичным процессу оценки содержательной компоненты, также предполагая выделение массивов токенов и n-грамм, и оценки их относительно присутствия в новостной единице. Однако, в случае тональной оценки, данные массивы должны отражать эмоциональный окрас новостной единицы. Таким образом, применение сформированной ранее методологии ограничено спецификой токенов, определяющих эмоциональный окрас текстовой информации. Необходимо отметить, что данная характеристика естественной цифровой информации определяется не только и не столько конкретными токенами, как их морфемной и лексической спецификой.

В связи с вышесказанным формирование автоматизированных инструментов оценки тональной компоненты основывается на принципе «обучения с учителем», с учетом использования в качестве зависимой переменной некого качественного параметра, отражающего эмоциональный окрас анализируемого объекта и оцениваемого эвристически. Одним из наиболее известных и эффективных инструментов оценки тональной компоненты текстовой информации, представленной на русском языке, является библиотека Dostoevsky. Данный инструмент позволяет идентифицировать распределение таких эмоциональных компонент текста как «уровень позитивности», «уровень негативности» и «уровень нейтральности». Безусловно, с точки зрения эмоционального окраса в новостной информации доминирует условная нейтральность. Однако аналитический интерес может представлять динамика относительного снижения нейтральности в корреляции с уровнем присутствия одного или нескольких выделенных тематических кластеров. Также, надо отметить потенциальную значимость программирования тонального отклика аудитории, за счет намеренного

повышения дисперсии эмоционального окраса относительно единовременно представленных новостных единиц. Данный подход сущностно является эмоциональной провокацией аудитории, однако в краткосрочном периоде подобная стратегия может значительно повысить потенциальный охват и вовлеченность потребителей информации. Детальный алгоритм дифференцированной оценки тональной компоненты извлеченных новостных единицах представлен на рисунке 8.

Полученные результаты позволяют сформировать первичный аналитический датафрейм, структурированный в соответствии с хронологическим распределением извлеченных новостных единиц, и являющийся завершением первичного логического этапа методологии системного анализа информационной среды, в разрезе новостной информации. Следующим шагом является анализ и математическая формализация сущности и структуры связей между оцененными квантификаторами извлеченной новостной информации, что будет рассмотрено в последующих статьях научной группы.

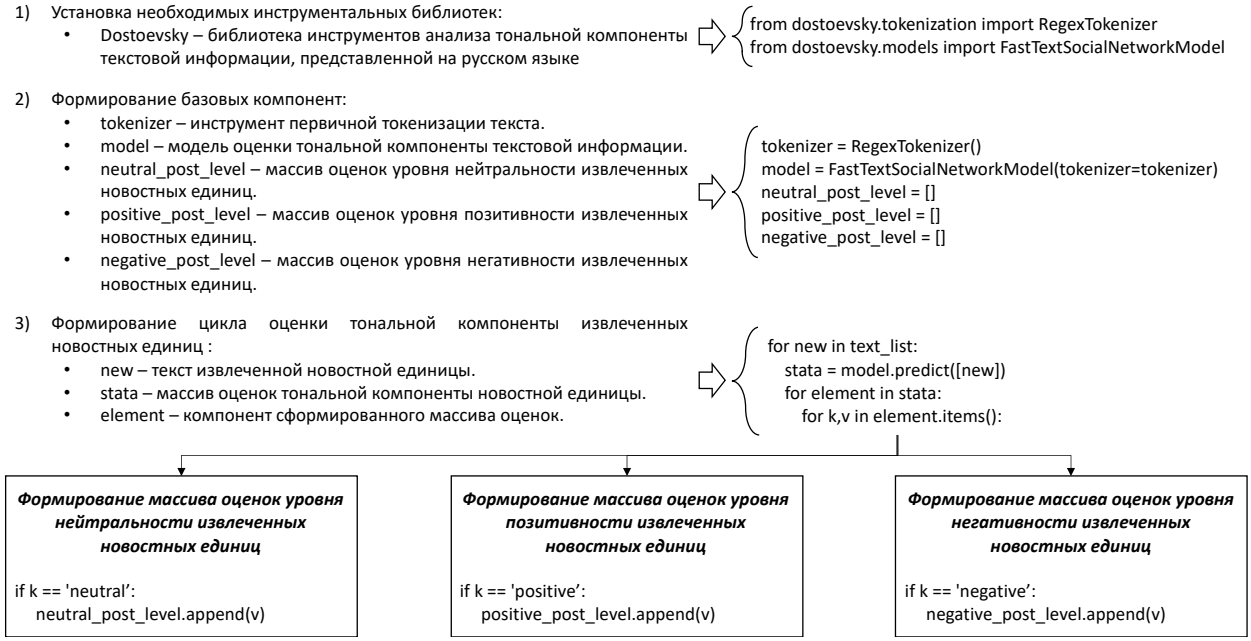

Рисунок 8. Алгоритм дифференцированной оценки тональной компоненты извлеченных новостных единицах

Библиографический список

1. E-commerce идет в гору.— Текст: электронный // Comnews: [сайт].— URL: https://www.comnews.ru/content/212828/2021–01–29/2021-w04/e-commerce-idet-goru (дата обращения: 14.02.2021).
2. Больше россиян в соцсетях сидят только японцы. Цифры.— Текст: электронный // C-News: [сайт].— URL: https://www.cnews.ru/news/top/2019–02–07_issledovanie_bolshe_rossiyan_v_sotssetyah_sidyat (дата обращения: 14.02.2021).
3. Сущность, функции и виды чувств и эмоций.— Текст: электронный // Энциклопедия Экономиста: [сайт].— URL: https://www.grandars.ru/college/psihologiya/emocii-i-chuvstva.html (дата обращения: 14.02.2021).
4. *Конников Е.А., Терентьева Д. А., Конникова О. А.* Анализ уровня устойчивого потребления в контексте цифровой информационной среды // Экономические науки. 2020. № 192. С. 114–125.
5. Термин: Лемматизация.— Текст: электронный // Система PromoPult: [сайт].— URL: https://promopult.ru/library/Лемматизация (дата обращения: 14.02.2021).
6. *Ершова А.В., Родионов Д. Г., Конников Е. А., Конникова О. А.* Системный анализ привлекательности банков для представителей ВИП-сегмента потребителей в рамках цифровой информационной среды // Экономические науки. 2021. № 194.

7. *Родионов Д.Г., Конников Е. А., Мугутдинов Р. М.* Системный анализ конкурентоспособности цифрового предприятия в рамках информационной среды // Экономические науки. 2020. № 193. С. 394–401.
8. *Конников Е.А., Вольвач О. С., Конникова О. А.* Маркетинговые решений в сфере альтернативной энергетики, основанные на формировании направленного информационного потока в цифровой среде // Экономические науки. 2020. № 193. С. 63–68.
9. *Родионов Д.Г., Конников Е. А., Алферьев Д. А.* Информационный капитал предприятия как целевой показатель развития в рамках цифровых экономических систем // Экономические науки. 2020. № 190. С. 131–137.