

Статистические выбросы как исследовательский объект в задачах социально-психологической диагностики*

Ю.М. Кузнецова, М.А. Станкевич, Н.В. Чудова

Федеральный исследовательский центр «Информатика и управление»,
Российской академии наук, г. Москва, Россия

Аннотация. Работа посвящена обсуждению способов интерпретации результатов анализа текстов сетевых дискуссий, представленных в виде статистически значимых отклонений («выбросов») анализируемых численных показателей от средних значений для нескольких лет наблюдения. Наличие выбросов предложено рассматривать в качестве маркера превышения естественной изменчивости диагностируемых показателей, что может свидетельствовать о влиянии на текстовую деятельность особых психологических состояний авторов, соотносимых с их реакцией на события различного рода, имеющие социальную значимость. Характеризующие реакцию участников сетевой коммуникации лексические и грамматические особенности текстов определялись с помощью инструмента анализа TITANIS на материале корпусов сетевых дискуссий с фиксированной тематикой и фоном. Предложено несколько способов представления и интерпретации паттернов выбросов анализируемых численных показателей в их статике и динамике, позволяющих составить представление о выраженности и эмоциональной направленности отношения коммуникантов к предмету сетевого обсуждения.

Ключевые слова: интеллектуальный анализ текста, сетевые дискуссии, социально значимые события, статистический выброс, эмоциональная реакция.

DOI: 10.14357/20790279250307 **EDN:** PFCGYP

Введение

Статистические методы давно и прочно заняли место главного инструмента в работе с данными психологического исследования. Круг таких понятий как «среднее значение», «медиана», «стандартное отклонение» соотносится в сознании исследователя с представлением о закономерности и потому усилия психологов обычно направлены на сбор максимально большого числа испытуемых и их реакций. При создании психодиагностических методик разработчики стараются довести объем наблюдений и их согласованность до такого уровня, когда картина «обычного», «среднего», «наиболее распространенного», «встречающегося почти у всех» реагирования испытуемых на тот или иной стимул стала бы совершенно ясной, а сильно отклоняющиеся случаи встречались бы редко и на представления о порождающем реакцию психологическом механизме не влияли бы.

Сама процедура создания психодиагностической методики включает в себя работу по ее стандартизации.

Все это совершенно естественно и правильно, однако, когда дело доходит уже до применения методики и измерения показателей в конкретном обследовании, случаи нетипичного, нестандартного реагирования оказываются за пределами исследовательского внимания. А между тем эти особые случаи, так называемые выбросы, могут быть важными источниками информации о происходящем в конкретной ситуации. На эту проблему еще век назад обращал внимание исследователей классик экспериментальной психологии Курт Левин. В своей в программной статье «Закон и эксперимент в психологии» (1927 г.) он писал: «Проведение экспериментов и анализ их результатов были в огромной степени подчинены духу статистики, основывающемуся на частоте повторения одного и того же опыта и вычислении средних значений. При такого рода статистическом подходе отдельные и даже многочисленные исключения играют весьма незначительную роль» [1, с. 26].

* Исследование выполнено в рамках научной программы Национального центра физики и математики (направление №9 «Искусственный интеллект и большие данные в технических, промышленных, природных и социальных системах»).

Рассматриваемая в работе проблема частично пересекается с вопросом выявления аномалий в социальных сетях. Актуальность данной проблемы подтверждается наличием заметного количества обзорных статей по теме [2-4]. И хотя нет одного строго устоявшегося определения, что считать аномалией, обычно речь в исследованиях по теме идет о выявлении поведения, которое значительно отличается от среднего поведения большинства участников социальной сети, которыми могут являться как отдельные пользователи, так и группы или сетевые сообщества [2]. Чаще всего такие исследования фокусируются на задачах выявления в социальной сети фейковых новостей, мошенничества, небезопасного контента и событий, вызывающих сильную реакцию у пользователей. Не редко данная задача рассматривается с точки зрения анализа графов и соответствующих методов выявления аномалий в графах, где в качестве вершин выступают отдельные пользователи [5-7].

С точки зрения обработки выбросов данных, примененный в текущей работе подход относится к тем случаям, где выбросы не являются чем-то, что мешает обработке данных, а наоборот, являются предметом интереса для исследования [8], что обусловлено природой анализируемых текстовых параметров текста. Например, аналогичный подход может применяться в задачах выявления мошенничества с банковскими картами или постановления медицинских диагнозов [9].

В данной работе мы хотим показать, как анализ случаев, характеризующихся при статистической обработке как выбросы, может быть встроен в интерпретацию данных и даже дать импульс для дальнейших практикоориентированных и фундаментальных исследований. Использование созданного в ФИЦ ИУ РАН инструмента интеллектуального анализа текстов для определения психологических и лингвистических показателей дискуссий, проходивших на протяжении ряда лет в так называемых городских сообществах социальной сети ВКонтакте, позволило определить роль анализа «необычных» дискуссий в задаче мониторинга социальной напряженности. Ниже показано, что «выбросы» могут служить маркерами всплесков социальной напряженности, т. е. само по себе появление реакций, резко выходящих за пределы среднестатистических, следует рассматривать как важный показатель в динамике социально-психологических процессов. Кроме того, дискуссии, оказавшиеся по некоторым важным психологическим показателям необычайно интенсивными, могут быть рассмотрены как «яркие представители» определенного типа обсуждений, т.е. «нестандарт-

ные» случаи могут служить основой для типологизации всего набора изучаемых объектов.

1. Материалы исследования

В качестве материала исследования были использованы тексты сетевых сообществ города с численностью населения до 100 тысяч в Нижегородской области. Используя встроенный поисковый механизм социальной сети ВКонтакте, мы отобрали самые активные сообщества, посвященные городским событиям и новостям данного города. При помощи официального API социальной сети были выгружены все дискуссии из данных сообществ за период с начала 2020 года по конец 2022. Под дискуссией, в рамках данной работы, понимается коллекция всех комментариев под одним постом. Всего было собрано порядка 83 тыс. постов и 660 тыс. комментариев к ним. Для дальнейшего рассмотрения были отобраны только те дискуссии, где в обсуждении участвовало минимум 10 уникальных пользователей. Объем такой выборки составил 10881 дискуссий, а суммарное количество комментариев в них составило 414224. Помимо этого, в данных были выделены дискуссии на медицинскую тематику. Для этого использовался подготовленный экспертами словарь медицинской лексики (объем 1010 слов, например, акушер, вирусный, вылечиться, грипп, кашель, реанимировать). Слова в тексте комментариев и постов были приведены к леммам при помощи инструмента морфологического анализа MyStem [10]. В ходе сопоставления лемм слов в комментариях и постах со словарем, была рассчитана их встречаемость, нормализованная относительно общего объема слов в дискуссии. Дискуссии со значением выше 70 процентиля показателя встречаемости слов по выборке были определены как относящиеся к медицинской тематике. Их количество составило 3346.

Все собранные комментарии были обработаны при помощи инструмента TITANIS. Общий функционал психодиагностического инструмента автоматического анализа текста TITANIS представлен в [11]. В данном исследовании анализ текстов проводился по параметрам, обеспечивающим выявление частоты представленности в текстах следующих характеристик: психолингвистических маркеров возбуждения (коэффициент Трейгера, коэффициент опредмеченности действия, коэффициент логической связности), эмотивных предикатов (позитивы, негативы, амбиваленты, деэмотивы; эмотивы базовых эмоций), направленности фрустрационного реагирования (экс-

трапунитивная, интропунитивная, импунитивная), тематических групп слов (ТГС), связанных с состоянием стресса (эмоциональный, рациональный и предметно-социальный аспекты), элементарных синтаксических характеристик, позволяющих проводить комплексный лингвистический анализ. В ходе обработки для каждого комментария было рассчитано 75 параметров. На следующем шаге показатели комментариев группировались относительно дискуссий, к которым они принадлежали, и рассчитывались средние значения параметров для дискуссий, далее понимаемые как показатели самих дискуссий. Полученные значения параметров и их распределения были использованы для расчета межквартильного размаха (IQR):

$$IQR = Q3 - Q1,$$

где $Q3$ – верхний квартиль значений параметра, $Q1$ – нижний квартиль значений параметра.

Выбросами принимались значения текстового параметра, которые соответствовали условию:

$$p > Q3 + IQR * 1.5,$$

где p – значение текстового параметра дискуссии.

Таким образом, рассматривались только выбросы с высокими значениями. Данная процедура выполнялась для каждого параметра отдельно. В итоге каждая дискуссия была описана в виде вектора признаков, где каждый элемент в бинарном виде отражает – находится ли значение того или иного параметра дискуссии в зоне выброса или нет.

В нашем случае «физический смысл» отобранных с помощью методов матстатистики данных можно охарактеризовать бытовым выражением «бурная дискуссия» – дискуссии, имеющие статистически значимые выбросы по измеряемым параметрам. Какие темы в сообщениях городских интернет-сообществ вызвали повышенную эмоциональную реакцию дискуссанта, в какой период времени интересующая исследователя тема выступала для участников дискуссий как стрессогенная, чему посвящены дискуссии, вызывающие повышенную агрессию их участников, и насколько единодушны в своем отношении к обсуждаемым предметам дискуссанта – на все эти вопросы можно попытаться получить ответы, анализируя данные о «бурных дискуссиях»

2. Результаты и их обсуждение

Анализ дискуссий с существенно повышенной эмоциональной вовлеченностью участников может вестись в двух направлениях – анализ динамики таких дискуссий с привязкой к социаль-

но значимым событиям в жизни города, региона, страны и анализ характерных для таких бурных дискуссий особенностей эмоционального реагирования дискуссанта.

Анализ динамики предполагает дополнительную процедуру в обработке данных. Если речь идет о реакции на происходящее событие или событие с долгими и значительными последствиями для участников сетевой коммуникации, необходимо перейти от рассмотрения данных по дням к анализу данных по неделям или даже месяцам.

В нашем случае корпус сетевых дискуссий содержал данные об обсуждениях, проходивших в период эпидемии коронавирусной инфекции и противоэпидемических мероприятий. Само событие «пандемия» имело длящийся характер и включало в себя ряд федеральных, региональных и городских событий, оказывающих серьезное влияние на жизнь людей. Соответственно, целесообразным представлялся анализ понедельной динамики эмоционального состояния пользователей соцсети. Можно было ожидать, что первые несколько недель эпидемии отразятся в сетевых дискуссиях повышенным числом бурных эмоциональных обсуждений медицинских тем. На рис. 1 видно, что это действительно было так – в изучаемых нами городских сообществах число дискуссий на медицинские темы с необычно высокими показателями эмоциональности было значительным на протяжении всего периода начала эпидемии. В 2020 г. общее число высокоэмоциональных дискуссий на медицинские темы резко увеличилось в середине марта, когда государством были введены первые противоэпидемические меры. Судя по нашим данным, острое стрессовое состояние у участников городских интернет-сообществ продержалось весь апрель и начало мая, а затем эмоциональность обсуждений медицинских тем стала постепенно снижаться, хотя до середины лета еще оставалась повышенной по сравнению с данными за 2021 и тем более 2022 годы.

Такой способ представления данных позволяет отразить выраженность тенденций, поскольку при этом выделяются пики и периоды наибольшего повышения количества выбросов для каждого года наблюдения, что может быть связано с более нестабильным состоянием коммуникантов.

Другой способ – сравнение тенденций – позволяет выявить специфику отношения к медицинским темам на фоне других предметов. На рис. 2 представлены данные по сравнению количества выбросов между годами для медицинских и немедицинских дискуссий. Можно видеть,

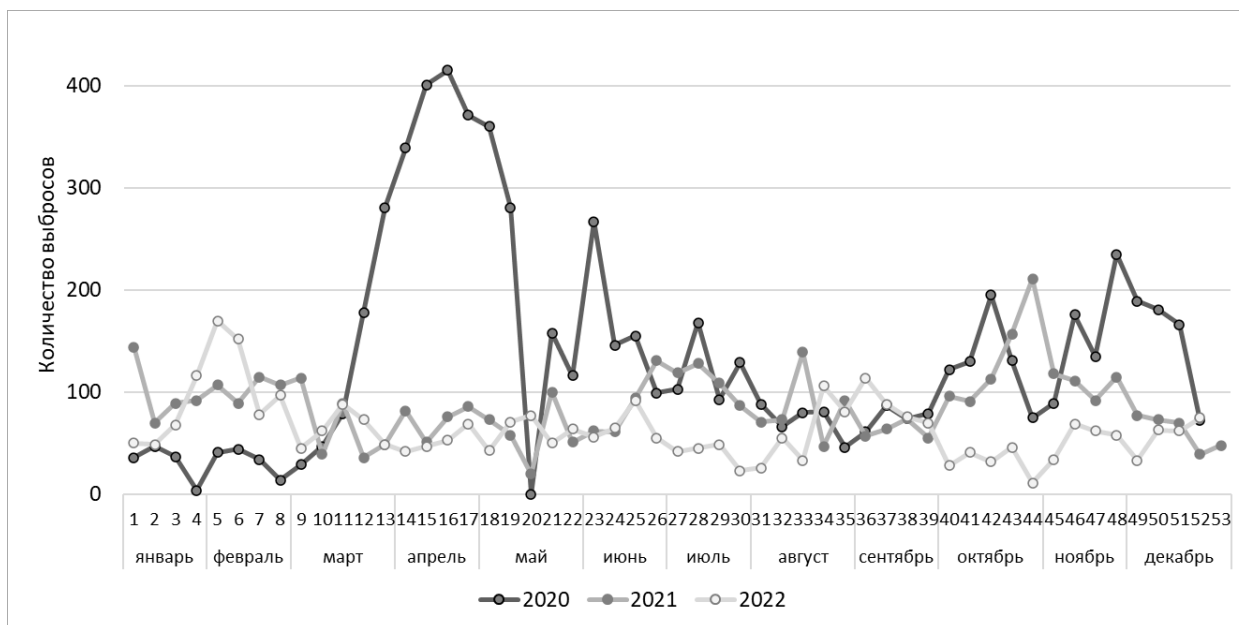


Рис. 1. Данные о количестве выбросов в дискуссиях на медицинские темы по всем показателям TITANIS в 2020–2022 гг. по неделям и месяцам

что более близкие значения количества выбросов фиксируются для 2020 г. С учетом того, что дискуссий на медицинские темы более чем в два раза меньше, чем фоновых, такое сближение может свидетельствовать о выраженной напряженности медицинской темы для коммуникантов. В 2022 г. годовые линии количества выбросов для дискуссий на медицинские и немедицинские темы сглаживаются, что может свидетельствовать о снижении общего эмоционального напряжения коммуникантов.

Третий способ направлен на выявление характера реакций населения. Приведенная на рис. 3 форма представления материала отражает общее количество выбросов определенного показателя TITANIS, а также вклад дискуссий на медицинские и немедицинские темы в динамику этого показателя. В данном случае нам была интересна динамика готовности граждан самостоятельно разрешать стоящие перед ними проблемы, в качестве текстового проявления которой выступает показатель i — один из маркеров TITANIS, характеризующих реакцию на фрустрацию авторов текстов. Как можно видеть, в самом начале эпидемии, в конце марта 2020 г., эта готовность в отношении проблем, связанных со здоровьем и здравоохранением, резко выросла и оказалась значительно выше значения этого показателя при обсуждении немедицинских тем.

Четвертая задача, которую можно решать с опорой на данные о выбросах, — выявление объек-

та, на который направлены или которые спровоцировали эмоциональные реакции пользователей в таких бурных дискуссиях. Так, на рис. 4 можно видеть, что пики частоты упоминания властных институтов и их представителей, т.е. максимальная вовлеченность «властей» в проблемы, связанные с медициной, фиксируется для весны 2020 г. — время принятия на федеральном уровне первых противоэпидемических мер в России. За другими пиками частотности упоминания «властей» в дискуссиях на медицинские темы также могут стоять иные события, связанные и с решениями руководства региона или города. Таким образом, этот способ представления данных позволяет проводить ретроспективный анализ отношения к действиям властей, если аналитик имеет возможность сопоставить график эмоционального реагирования при упоминании «властей» и линейку событий местного, регионального и федерального масштабов.

Наконец, данные о выбросах позволяют сравнивать уровень эмоциональной значимости определенных сфер социальной реальности, упоминаемых коммуникантами в сетевых дискуссиях. Так, сравнение частотности тематической лексики, отраженной на рис. 5, позволяет обнаружить пики и периоды наибольшей представленности лексики из ТГС «Власти» и «ЖКХ». Предметы конкретных обсуждений, если это предусмотрено задачами исследования, могут быть выявлены путем ручного ана-

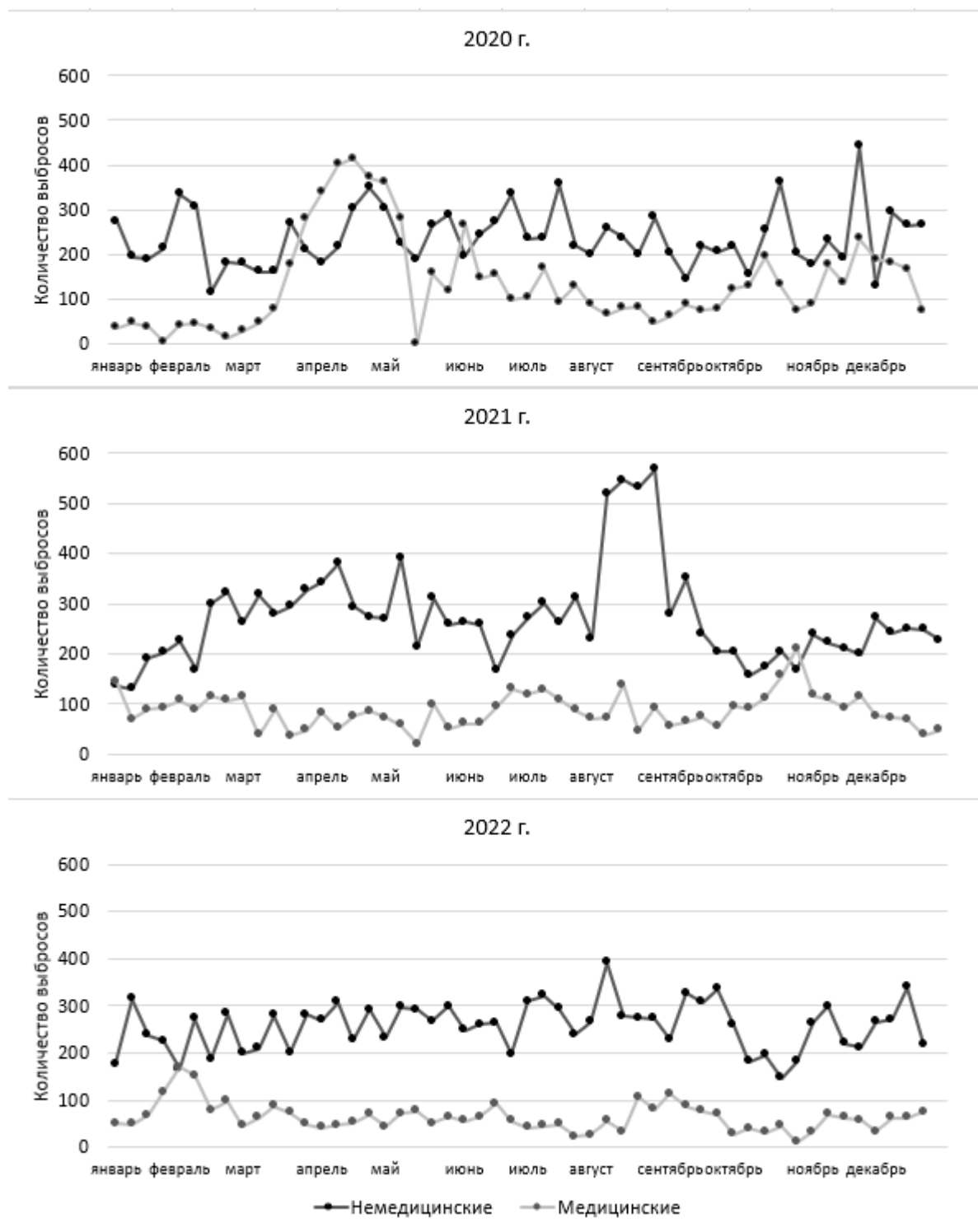


Рис. 2. Данные о количестве выбросов по всем показателям TITANIS в 2020–2022 гг. в дискуссиях на медицинские и немедицинские темы по месяцам

лиза текстов этих дискуссий. Также при сравнении частотности лексики из разных ТГС можно оценить относительную значимость соответствующих тем обсуждения. Если в 2020 г. лек-

сика ТГС «Власти» более частотна, чем лексика ТГС «ЖКХ», то к 2022 г. разрыв сокращается за счет снижения частотности лексики ТГС «Власти». Таким образом, в анализируемых сетевых

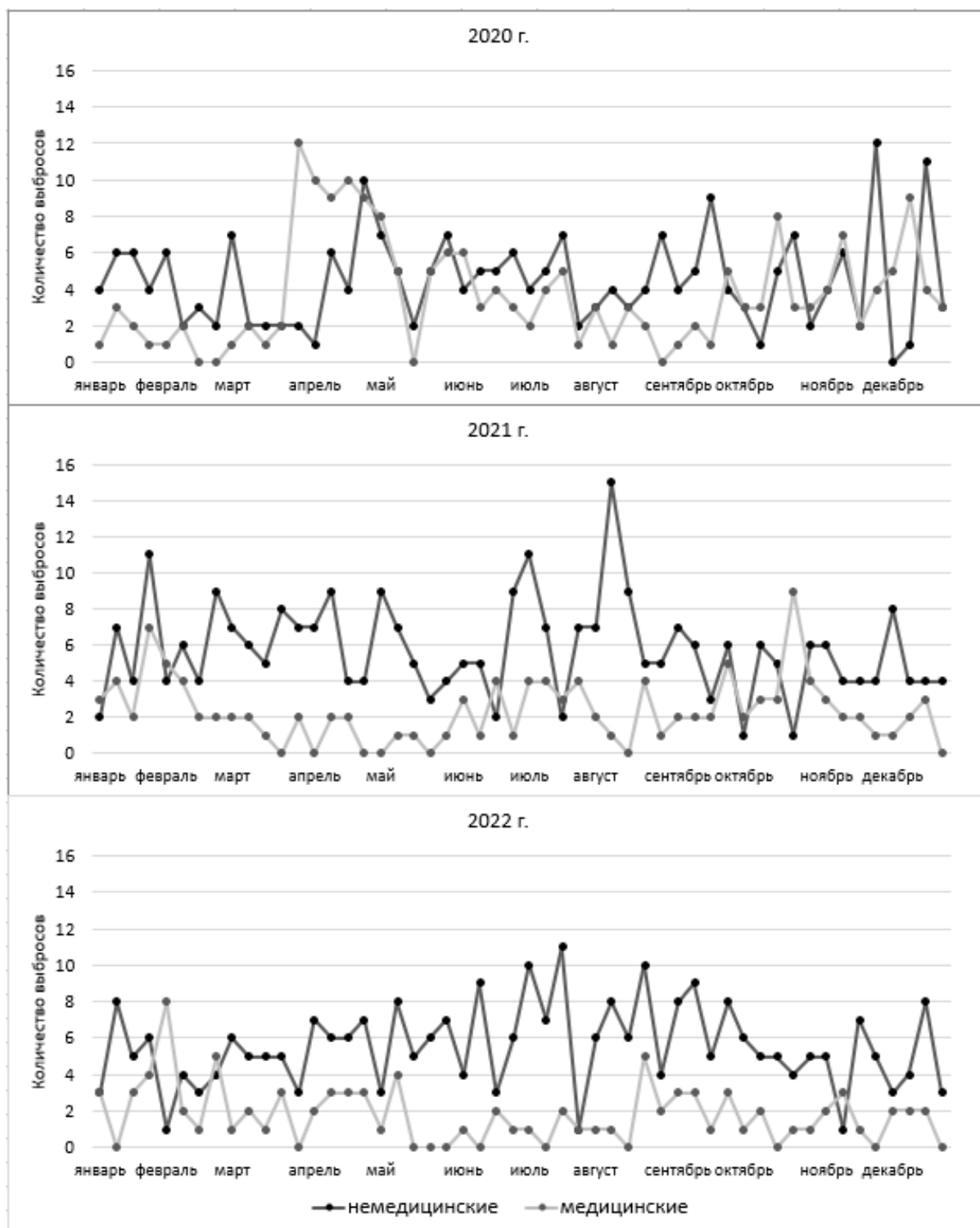


Рис. 3. Данные о количестве выбросов по показателю «i» TITANIS в 2020–2022 гг. в дискуссиях на медицинские и немедицинские темы по месяцам

дискуссиях количество упоминаний конкретных управленческих решений, самих властных институтов и их представителей постепенно снижается, в то время как различные аспекты

коммунального хозяйства и инфраструктуры представляют собой предмет стабильного интереса коммуникантов, что наглядно демонстрирует рис. 6.

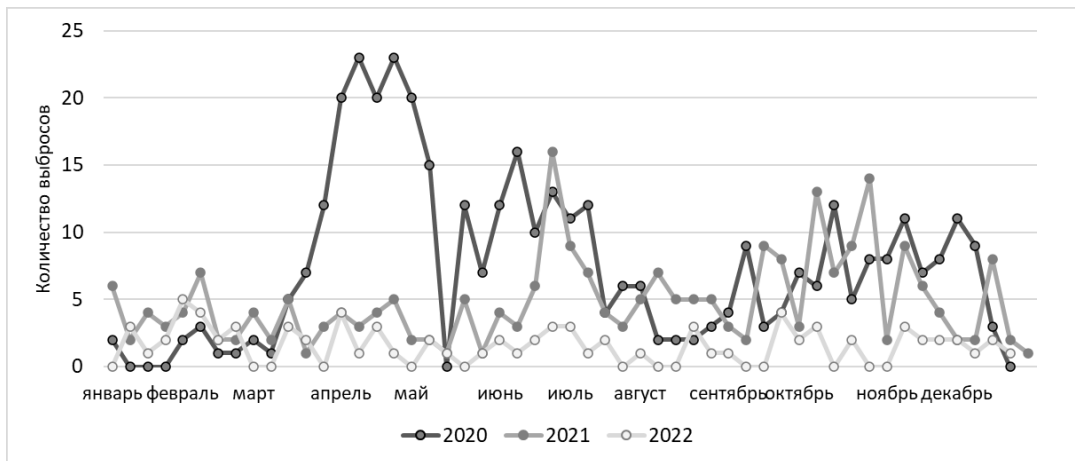


Рис. 4. Данные о количестве выбросов частотности лексики из ТГС «Власти» в 2020–2022 гг. по месяцам

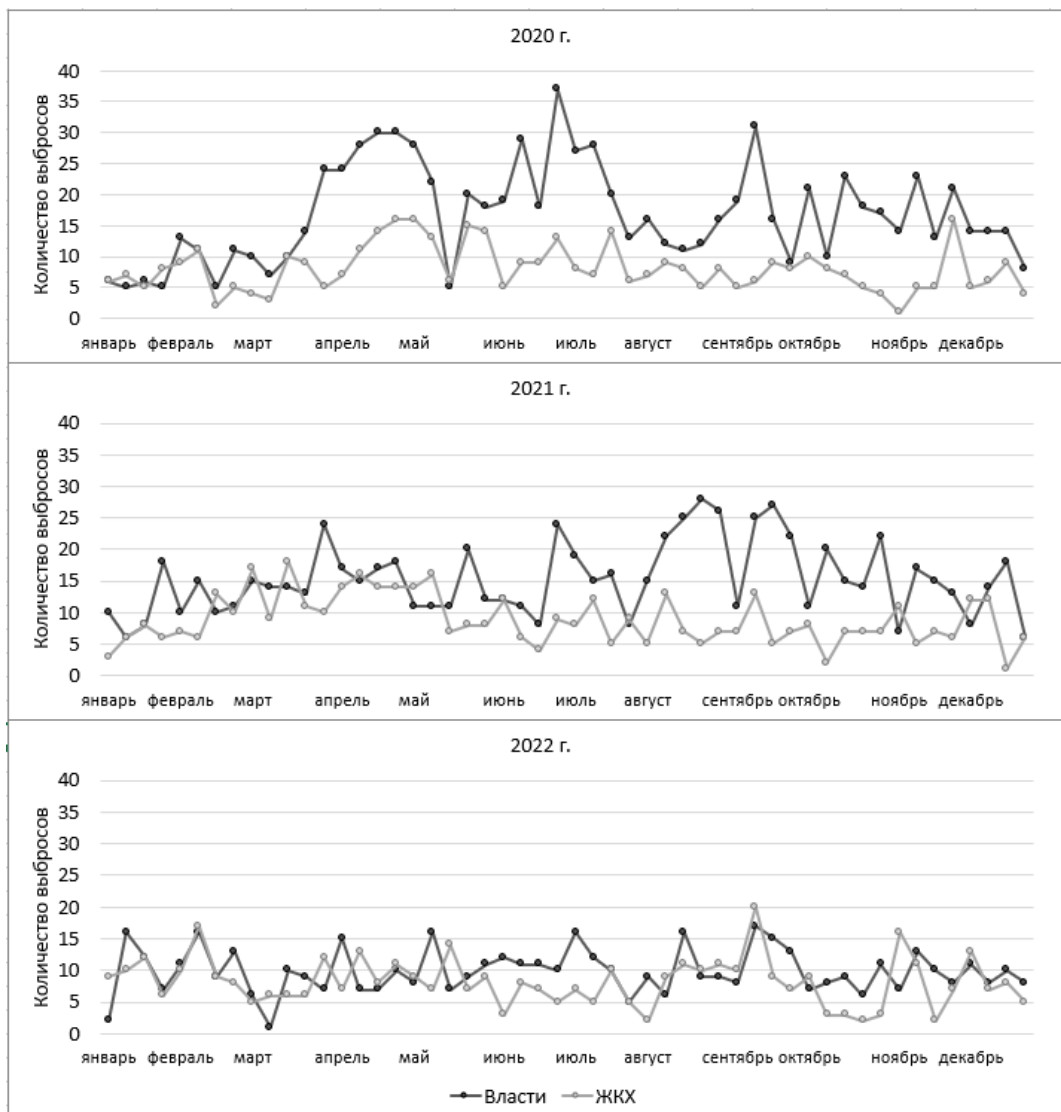


Рис. 5. Данные о количестве выбросов частотности ТГС «Власти» и «ЖКХ» в 2020–2022 гг. по месяцам

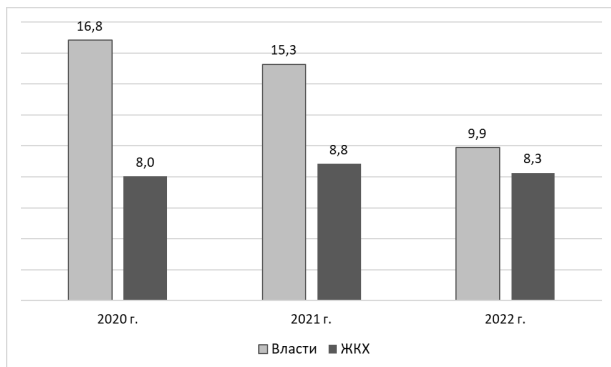


Рис. 6. Среднегодовое количество выбросов частотности лексики из ТГС «Власти» и «ЖКХ»

Второе направление – анализ целостной картины эмоционального реагирования дискуссанта, характерной для дискуссий с существенно повышенными показателями, предполагает работу по выделению типов дискуссий. Типологизация сетевых обсуждений может проводиться по самым разным основаниям и, как правило, их выбор определяется в первую очередь интересами исследователя и его представлениями о существенном, важном при изучении данного объекта. Метод анализа выбросов позволяет сместить акценты в работе по типологизации, так что реальность конкретных дискуссий начинает приобретать больший вес в анализе. Такой «индуктивный» путь для социо-гуманитарных дисциплин сейчас не очень характерен, хотя, например, самые интересные типологии в области психологии состояний и психологии характера созданы именно таким, идущим от наблюдений за яркими случаями, способом.

При типологизации сетевых дискуссий методом анализа выбросов существенную роль будет играть сам собранный корпус текстов. Так, материалом для анализа могут служить обсуждения, проходящие на разных интернет-площадках – в блогах лидеров общественного мнения, в сообществах, созданных для обсуждения определенных тем, на сайтах СМИ, политических партий или общественных движений и др. Материал может собираться и по тематическому принципу, когда исследователь отбирает дискуссии, посвященные одной теме, из разных социальных сетей и каналов. Сбор материала может проводиться и по принципу временного среза, когда корпус составляется из дискуссий, прошедших в какой-то один, значимый для жизни общества, период. Возможен и интерес исследователя к отдельному, выбранному им для наблюдения, объекту – к конкретному городу, к определенному течению в подростковой субкультуре, к заданному профессиональному или

соседскому сообществу и т.д. В мировой науке на данном этапе изучения сетевых сообществ нет достаточного материала для построения общей типологии сетевых дискуссий, поэтому метод анализа выбросов представляется релевантным степени разнообразия самих сетевых дискуссий.

В нашем корпусе дискуссий, проходящих в новостных сетевых сообществах небольшого приволжского города, метод анализа выбросов, включающий в себя статистическую оценку значений измеряемых параметров и психологическую интерпретацию системы параметров с повышенными значениями, позволил выделить три типа дискуссий обвинительного характера: спор, единодушное недовольство, обсуждение животрепещущей темы. Такие дискуссии отличаются высоким уровнем выражаемого недовольства действиями / высказываниями собеседников или третьих лиц / институций / институтов. В психодиагностике такие тексты считаются проявлением так называемой экстрапунитивной реакции на фрустрацию с фиксацией на защите «Я», в нашем исследовании они выявлялись автоматически с помощью специального модуля TITANIS. Ниже приведены данные по дискуссиям, отличающимся в нашем корпусе высоким уровнем обвинений.

Первый тип обвинительных дискуссий (1) характеризуется двойственным отношением к обсуждаемому объекту. Так, в дискуссии, прошедшей 17 декабря 2022 года, в зону выбросов попали значения следующих параметров: ТГС «ЖКХ»; ТГС «Рациональная оценка положительная»; ТГС «Рациональная оценка отрицательная»; ТГС «Социальность положительная»; ТГС «Социальность отрицательная»; ТГС «Инвективы» и ТГС «Мягкие инвективы»; Эмотивы-позитивы и Эмотивы Радость; Эмотивы Удивление; ТГС «Когнитивная деятельность и коммуникация». Как можно видеть, собеседники демонстрируют как положительную, так и отрицательную оценку обсуждаемого объекта; проявляют как солидарность, так и стремление к размежеванию. В дискуссии встречается как много предикатов с семантикой положительных эмоций, так и много ругательств. При этом значительно повышена доля лексики коммуникации и часто выражается удивлением. Таким образом, такого рода дискуссию можно охарактеризовать как горячий *спор* по поводу обсуждаемого объекта (в данном случае объектом выступает работа коммунальных служб города).

Второй тип дискуссий, отличающихся высоким уровнем агрессии, обвинительных реакций, – *единодушное недовольство* объектом. В нашем корпусе таким объектом оказались власти (руко-

водство отдельными организациями и в целом горадминистрация – (2.1.).

В дискуссии, прошедшей 31 октября 2020 года, в зону выбросов попали значения следующих параметров: ТГС «Власти»; Доля глаголов 3 лица; ТГС «Деструктивная активность»; ТГС «Аффектогенная лексика»; ТГС «Социальность отрицательная»; ТГС «Обсценная лексика»; ТГС «Мягкие инвективы»; Эмотивы-негативы; Эмотивы Печаль; Эмотивы Покой и ТГС «Рациональная оценка положительная». В зоне выбросов оказались и значения такого параметра, как Сообщение-информирование. Система параметров, чьи значения в данной дискуссии оказались в зоне выбросов, характеризует всеобщее негативное отношение к предмету обсуждения. Это выражается в частом употреблении лексики деструкции, аффектогенности, размежевания, в использовании значительного числа обсценцизмов и инвектив, в фиксации своих негативных переживаний. Судя по высокой доле глаголов третьего лица речь в дискуссии идет о чьих-то действиях (бездействии), которые и вызывают недовольство. А так как, в отличие от дискуссии первого типа (1), здесь мы не видим повышения лексики коммуникации, разумным будет предположение о том, что осуждению подвергаются не собеседники, а третьи лица. Необходимо отметить, что при всем при этом наблюдается и повышенное число реплик, содержащих не авторскую реакцию, а информацию, взятую из внешних источников, – это может быть информация из СМИ или справочные сведения или даже сообщения, распространяемые модератором-представителем горадминистрации. Можно предположить, что и повышенная встречаемость лексики положительной рациональной оценки и предикатов покоя характерны именно для этих информирующих реплик.

Разумеется, для более точного определения происходящего в конкретной дискуссии необходима ручная проверка всего текста, но, тем не менее, этот пример показывает, что метод анализа выбросов позволяет выдвигать обоснованную гипотезу для последующей работы с материалом уже самих управленцев/модераторов.

Другой пример дискуссий, содержащей признаки единодушного недовольства – дискуссия, прошедшая 14 ноября 2020 года. Для этой дискуссии в зоне выбросов оказались значения следующих параметров: ТГС «Власти»; ТГС «Рациональная оценка отрицательная»; ТГС «Деструктивная активность»; ТГС «Инвективы»; Эмотивы Печаль; Эмотивы-амбиваленты; Эмотивы Удивление; «Коэффициент Трейгера; Коэффициент опредмеченности действия. В этой дискуссии наблюдается

опять общее для всех участников осуждение властей, усиленное еще и эмоциональным напряжением, о чем свидетельствует повышение обоих клинических психолингвистических показателей (коэффициентов) глагольности. При этом никаких признаков активных попыток поделить информацией, поданной в нейтральном или позитивном ключе, здесь уже не обнаруживается.

Наконец, метод анализа выбросов позволил в медицинском подкорпусе нашего корпуса выделить такой тип дискуссий, в котором обвинительные реакции возникают в рамках общей эмоциональной вовлеченности дискуссанта в тему обсуждения. Дело в том, что тема медицины сама по себе является эмоциогенной, так что если работу ЖКХ, или действия властей, или, например, природные катастрофы в других регионах жители могут обсуждать с разной степенью вовлеченности/отстраненности, то все, что связано со здоровьем, заставляет человека переживать и, в том числе, много говорить о своих и чужих эмоциях. Такой тип дискуссии и предлагается обозначать как *обсуждение трудных тем*. Как и в случае с предыдущим типом дискуссий, здесь наблюдаются вариации обсуждения.

Для первого подтипа (3.1) характерен значительный объем информирующих высказываний и сосредоточенность на проблемах здоровья, своего и родственников. Так, в дискуссии прошедшей 21 декабря 2022 года, в зоне выбросов оказались значения следующих параметров: ТГС «Эмоции и телесные состояния негативные пассивные»; ТГС «Социальность положительная»; ТГС «Рациональная оценка отрицательная»; ТГС «Мотивация, деятельность и напряжение»; Сообщение-информирование; Эмотивы-негативы; Эмотивы Страх; Эмотивы Печаль; Эмотивы-деэмотивы; Эмотивы Покой; Число отрицательных приставок и форм / Число слов.

В дискуссии второго подтипа (3.2) собеседники обсуждают скорее проблемы здравоохранения и, соответственно, значительное внимание уделяют работе управляющих органов и экономическим аспектам этой сферы. Так, в дискуссии прошедшей 15 ноября 2021 года, в зоне выбросов оказались значения следующих параметров: ТГС «Здравоохранение, демография и экология»; ТГС «Власти»; ТГС «Экономика»; ТГС «Рациональная оценка отрицательная»; ТГС «Деструктивная активность»; Эмотивы-деэмотивы; Эмотивы Радость; Эмотивы-позитивы.

Отметим в завершение, что практическая значимость получаемой методом анализа выбросов типологии дискуссий обусловлена тем, что,

разные типы дискуссий требуют разного участия модератора. В нашем случае это представитель администрации города и для него важно понимать каков должен быть объем, тональность и направленность информационных интервенций в обсуждение городских проблем. Проведенный анализ показывает, что любая дискуссия, типологически сходная с «бурными дискуссиями» (1) или (3.1) не требует вмешательства – собеседники заняты спором или обсуждением трудной темы и им важно лишь иметь площадку для канализации своих эмоций. При появлении дискуссий похожих на (2.2) или (3.2) требуется информирование подписчиков сообществ о мерах, предпринимаемых властями. Если же появляется дискуссия, сходная с (2.1), от модератора требуется внесение в ленту сообщества информации об оптимальном с точки зрения администрации города сценарии поведения горожан – это позволит как минимум канализировать социальное напряжение, а при удачной модерации будет способствовать и бесконфликтному разрешению обсуждаемой городской проблемы.

Заключение

Полученные в ходе исследования данные свидетельствуют о перспективности использования средств интеллектуального анализа текста на таком легкодоступном материале, как сетевые дискуссии в открытых сообществах в целях определения количественных (выраженность, продолжительность) и качественных (содержание, направленность) характеристик реакций населения на социально значимые события. Предложенный метод обработки собираемых в течение достаточно продолжительных промежутков времени данных текстового анализа сетевого контента демонстрирует чувствительность к реально наблюдавшимся в период наблюдения событиям, что может характеризовать его эмпирическую валидность как удовлетворительную и достаточную для применения в практических целях – таких, как организация информационной поддержки принятия управленческих решений и административного планирования. Развитие предложенного метода подразумевает дальнейшую работу по формированию эмпирически и теоретически обоснованной типологии отношения участников сетевого общения к различным явлениям и аспектам социальной жизни.

Литература

1. Левин К. Динамическая психология: избранные труды. 2001.
2. Xing L. et al. A survey on social network's anomalous behavior detection //Complex & Intelligent Systems. 2024;10(4):5917-5932.
3. Lunawat S., Rao J., Patil P. A Comprehensive Survey on Anomaly Detection in Social Media Networks: Challenges, Methods, and Future Directions //2024 4th International Conference on Sustainable Expert Systems (ICSSES). IEEE. 2024. P. 363-370.
4. Alam S., Faisal M. A comprehensive review: Anomaly detection techniques on social networking and its applications //Advances in Science, Engineering and Technology. 2025. P. 303-309.
5. Khan W. An exhaustive review on state-of-the-art techniques for anomaly detection on attributed networks //Turkish Journal of Computer and Mathematics Education (TURCOMAT). 2021;12(10):6707-6722.
6. Hassanzadeh R., Nayak R., Stebila D. Analyzing the effectiveness of graph metrics for anomaly detection in online social networks //International conference on web information systems engineering. – Berlin, Heidelberg : Springer Berlin Heidelberg. 2012. P. 624-630.
7. Anand K., Kumar J., Anand K. Anomaly detection in online social network: A survey //2017 International conference on inventive communication and computational technologies (ICICCT). IEEE. 2017. P. 456-459.
8. Blázquez-García A. et al. A review on outlier/anomaly detection in time series data //ACM computing surveys (CSUR). 2021;54(3):1-33.
9. Aggarwal C.C. An introduction to outlier analysis //Outlier analysis. Cham: Springer International Publishing. 2016. P. 1-34.
10. MyStem. URL: <https://yandex.ru/dev/mystem> [Дата обращения 07.08.2025].
11. Smirnov I., Stankevich M., Kuznetsova Y., Suvorova M., Larionov D., Nikitina E., Savelov M., Grigoriev O. TITANIS: A Tool for Intelligent Text Analysis in Social Media //Artificial Intelligence: 19th Russian Conference, RCAI 2021, Taganrog, Russia, October 11–16, 2021, Proceedings 19. – Springer International Publishing. 2021. P. 232-247. https://doi.org/10.1007/978-3-030-86855-0_16

Кузнецова Юлия Михайловна. Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия. Старший научный сотрудник. Кандидат психологических наук. Область научных интересов: психолингвистика, психосемантика, сетевая психодиагностика, компьютерный анализ текста. E-mail: kuzjum@yandex.ru

Станкевич Максим Алексеевич. Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия. Младший научный сотрудник. Область научных интересов: машинное обучение, обработка естественного языка, интеллектуальный анализ данных. E-mail: stankevich@isa.ru (Ответственный за переписку)

Чудова Наталья Владимировна. Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия. Старший научный сотрудник. Кандидат психологических наук. Области научных интересов: психология агрессии, психология интернета, когнитивное моделирование, сетевая психодиагностика. E-mail: nchudova@gmail.com

Statistical outliers as a research object in the tasks of socio-psychological diagnostics

Yu.M. Kuznetsova, M.A. Stankevich, N.V. Chudova

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia

Abstract. The paper presents a way to assess the reaction to socially significant events in online discussions, based on statistically significant deviations ("outliers") of numerical text indicators for some year's observations. We consider outliers to be a marker of exceeding the natural variability of the diagnosed indicators, that is, an evidence authors' psychological states correlated with the reaction to socially significant events affect textual activity. The TITANIS lexical and grammatical indicators of users' reactions were evaluated in the thematic and background corpus of online discussions. Some ways of representation and interpretation are proposed for the patterns of statics and dynamics of outliers, thereby the severity and emotional orientation of users' attitudes towards the topic of network discussion can be described.

Keywords: *text mining, online discussion, socially significant event, statistical outlier, emotional reaction.*

DOI: 10.14357/20790279250307 **EDN:** PFCGYP

References

1. Levin K. Dynamic psychology: selected works. 2001. (In Russ).
2. Xing L. et al. A survey on social network's anomalous behavior detection // Complex & Intelligent Systems. 2024;10(4): 5917-5932.
3. Lunawat S., Rao J., Patil P. A Comprehensive Survey on Anomaly Detection in Social Media Networks: Challenges, Methods, and Future Directions //2024 4th International Conference on Sustainable Expert Systems (ICSSES). IEEE. 2024. P. 363-370.
4. Alam S., Faisal M. A comprehensive review: Anomaly detection techniques on social networking and its applications //Advances in Science, Engineering and Technology. 2025. P. 303-309.
5. Khan W. An exhaustive review on state-of-the-art techniques for anomaly detection on attributed networks //Turkish Journal of Computer and Mathematics Education (TURCOMAT). 2021;12(10): 6707-6722.
6. Hassanzadeh R., Nayak R., Stebila D. Analyzing the effectiveness of graph metrics for anomaly detection in online social networks //International conference on web information systems engineering. – Berlin, Heidelberg : Springer Berlin Heidelberg. 2012. P. 624-630.
7. Anand K., Kumar J., Anand K. Anomaly detection in online social network: A survey //2017 International conference on inventive communication and computational technologies (ICICCT). IEEE. 2017. P. 456-459.
8. Blázquez-García A. et al. A review on outlier/anomaly detection in time series data //ACM computing surveys (CSUR). 2021;54(3):1-33.
9. Aggarwal C.C. An introduction to outlier analysis //Outlier analysis. Cham: Springer International Publishing. 2016. P. 1-34.

10. MyStem. URL: <https://yandex.ru/dev/mystem> [Дата обращения 07.08.2025].
11. *Smirnov I., Stankevich M., Kuznetsova Y., Suvorova M., Larionov D., Nikitina E., Savelov M., Grigoriev O.* TITANIS: A Tool for Intelligent Text Analysis in Social Media //Artificial Intelligence: 19th Russian Conference, RCAI 2021, Taganrog, Russia, October 11–16, 2021, Proceedings 19. – Springer International Publishing. 2021. P. 232-247. https://doi.org/10.1007/978-3-030-86855-0_16

Yulia M. Kuznetsova. Senior researcher, Ph.D. in psychology. Federal Research Center Computer Science and Control of Russian Academy of Sciences (FRC CSC RAS), Moscow, Russia. Research areas: psycho-linguistics, psycho-semantics, network psycho-diagnostics, text mining. E-mail: kuzjum@yandex.ru

Maksim A. Stankevich. Junior researcher. Federal State Institution “Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences”, Moscow, Russia. Research areas: machine learning, natural language processing, data mining, data analytics. E-mail: stankevich@isa.ru

Natalia V. Chudova. Senior researcher, PhD in Psychology. Federal State Institution “Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences”, Moscow, Russia. Research areas: psychology of aggression, psychology of the Internet, cognitive modeling, network psychodiagnostics. E-mail: nchudova@gmail.com