

Фильтрация сообщений в социальных сетях для выявления признаков чрезвычайных ситуаций по регионам

Я. А. Бекенева✉, И. А. Байдунов

Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В. И. Ульянова (Ленина), Санкт-Петербург, Россия

✉yabekeneva@etu.ru

Аннотация. Социальные сети уже достаточно давно используются не только в развлекательных, но и в информативных целях, в частности существует немало примеров сообществ, посвященных информированию о чрезвычайных ситуациях (например, ДТП) или обсуждению различного рода нештатных ситуаций в отдельно взятом городе или районе. В статье описывается подход к обработке сообщений из социальных сетей для выявления разного рода чрезвычайных ситуаций и определения числа людей, нуждающихся в помощи. Предлагается осуществлять поиск и фильтрацию сообщений по ключевым словам и заданным пользователями меткам. Представлен результат работы программы, реализующей предложенный подход, позволяющий обработать большие наборы данных и получить выходной файл, в котором наглядно отображаются требуемые виды помощи и ее объем. Результаты могут быть использованы в любой сфере жизнеобеспечения и для реагирования на нештатные ситуации.

Ключевые слова: социальные сети, сообщения, атрибуты сообщений, фильтрация, ключевые слова, метки, чрезвычайные ситуации

Для цитирования: Бекенева Я. А., Байдунов И. А. Фильтрация сообщений в социальных сетях для выявления признаков чрезвычайных ситуаций по регионам // Изв. СПбГЭТУ «ЛЭТИ». 2022. Т. 15, № 8. С. 50–56. doi: 10.32603/2071-8985-2022-15-8-50-56.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Original article

Filtering Social Media Posts to Identify Signs of Emergencies by Region

Ya. A. Bekeneva✉, I. A. Baydurov

Saint Petersburg Electrotechnical University, Saint Petersburg, Russia

✉yabekeneva@etu.ru

Abstract. Social networks have long been used not only for entertainment, but also for informative purposes, in particular, there are many examples of communities dedicated to informing about emergencies (for example, traffic accidents) or discussing various kinds of emergency situations in a single city or region. The article describes an approach to processing messages from social networks to identify various kinds of emergencies and determine the number of people in need of help. It is proposed to search and filter messages by keywords and user-specified labels. The result of the work of the program that implements the proposed approach is presented, which allows processing large data sets and obtaining an output file that clearly displays the required types of assistance and its volume. The results can be used in any area of life support and to respond to emergency situations.

Keywords: social networks, messages, message attributes, filtering, keywords, tags, emergencies

For citation: Bekeneva Ya. A., Baydurov I. A. Filtering Social Media Posts to Identify Signs of Emergencies by Region // LETI Transactions on Electrical Engineering & Computer Science. 2022. Vol. 15, no. 8. P. 50–56. doi: 10.32603/2071-8985-2022-15-8-50-56.

Conflict of interest. The authors declare no conflicts of interest.

Введение. Социальные сети – это интернет-площадки для общения, обмена информацией и контентом, прочих социальных взаимодействий. Они используются для работы, отдыха и развлечений, позволяют координировать между собой группы людей и имеют широкий набор функций. Соцсетями пользуются миллионы людей по всему миру, что делает их прекрасным инструментом интернет-маркетинга [1]. Существует множество различных социальных сетей, которые отличаются функциями и направленностью.

На основе полученных данных – результатов исследований социальных сетей, можно формировать совокупность свойств, присущих отдельному пользователю, и делать выводы о его личности и предпочтениях [2]. Также можно получить информацию о настроении группы пользователей в целом, выявить нежелательные ситуации при самоорганизации общества, что, в свою очередь, играет важную роль в решении многих экономических и социальных проблем [3]–[5]. Некоторые сведения могут быть полезны в обеспечении безопасности.

Постановка задачи. В результате исследования существующих методов анализа данных в социальных сетях было установлено, что исходные данные не могут быть обработаны через доступные сервисы так, как это требуется, а значит, для данных такого рода требуется создание специального приложения, которое бы выполняло все необходимые функции. Была поставлена задача создания метода обработки информации из сообществ в социальных сетях для их дальнейшего анализа. Результаты анализа общения жителей о чрезвычайной ситуации помогут сделать некоторые выводы для экстренного реагирования социальных служб.

Анализ сообщений в социальных сетях опирается на поставленную задачу. Зачастую целями анализа информации в социальной сети становятся: анализ аудитории для таргетированной рекламы [6], сбор данных для социологических исследований [7], поиск потенциально опасных для общества заявлений пользователей [8]. Методы анализа, используемые в данной отрасли, могут быть применены не только для определения товаров и услуг, интересных пользователю, но и в обеспечении безопасности целых регионов. Собрав данные из сети и проанализировав их, можно узнать, в каком регионе какая помощь может понадобиться.

В силу приватности и политики конфиденциальности социальных сетей анализировать можно

только открытые и доступные данные. К закрытым данным относится в основном личная переписка пользователей или данные в частных сегментах сети. Если изначально общение не приватное, как, к примеру, общение между людьми в открытом чате, в котором каждый пользователь может прочитать сообщение другого пользователя и выразить свое мнение, то данные из этого чата можно использовать для обработки и анализа, не нарушая при этом существующие нормы и правила (если не наложены дополнительные ограничения на использование информации).

Основная часть. Сообщения в социальных сетях характеризуются так называемыми атрибутами – необходимыми для анализа и структуризации свойствами сообщений. В качестве основных атрибутов выступают имя отправителя сообщения и текст сообщения. Любое сообщение, размещенное в социальной сети, обладает следующими атрибутами [9]:

1. *Дата и время.* Атрибут сообщения, благодаря которому можно выяснить время отправки сообщения.

2. *Имя пользователя.* Эту характеристику также называют аккаунтом, она входит в набор стандартных, так как несет в себе информацию о том, какой пользователь (аккаунт) отправил сообщение.

3. *Локация.* Данная характеристика присуща не всем сообщениям, но во многих социальных сетях при желании ее можно настроить. Она позволяет получить информацию о местонахождении пользователя, который отправил сообщение. Чаще всего это – географическая метка на карте или координаты. Существуют различные способы определить местоположение пользователя, но самым используемым остается метод определения геолокации устройства пользователя по IP-адресу. Этот способ не обеспечивает высокую точность, но определить местоположение с точностью до региона или города он может. Данный способ геолокации не требует взаимодействия с пользователем, кроме разрешения пользователя на получение этих сведений приложением. Обычно это регулируется в настройках или запрашивается самим приложением.

4. *Содержание сообщения.* Само сообщение служит одним из важнейших источников данных, ведь оно содержит основную массу информации, предоставленной пользователем в виде текста, фотографий или видео (в статье рассматриваются текстовые сообщения). При анализе информации, со-

держатся в сообщении, важны некоторые ключевые слова, которые используются прежде всего для поиска необходимой информации в тексте.

По данным атрибутам сообщения из социальной сети можно извлечь и систематизировать, например в таблицу, которая станет большой базой данных, данные в которой будут иметь структурированный вид. Легче всего выгрузить в таблицу сообщения за конкретный указанный временной период, так как это будет занимать намного меньше времени, а также эффективней обрабатываться, так как для сбора всего объема сообщений из социальной сети потребуется огромное количество памяти, а следовательно, и высокая мощность и производительность техники.

Смоделируем ситуацию, где существуют группы и сообщества в социальной сети, в которых интерес людей направлен не на развлечения и бизнес, а на четко выраженную цель – получение сведений для обеспечения безопасности в городе. При помощи обмена информацией можно подготовиться ко многим ситуациям, и при обострении серьезности положения иметь возможность отреагировать. В этом случае социальная сеть будет носить не просто развлекательный характер, она обретет огромную значимость для общества и многих сфер его жизни.

Предположим, сообщество касается конкретной области, где метеорологические службы объявили о некотором опасном природном явлении. Это могут быть как землетрясения, наводнения, цунами, смерчи, ураганы, песчаные бури, лавины, тропические циклоны, оползни, так и различная вулканическая активность. Стоит отметить, что метеорологические службы хотя и выдают прогнозы, не могут гарантировать их точность, так как природные явления могут вести себя непредсказуемо. К тому же ситуация всегда может измениться, поэтому большая территория в области предполагаемого природного явления находится в некоторой зоне опасности. Безусловно, мирное население всегда старается оградить и эвакуировать из опасной зоны заранее, но случается и так, что предсказываются небольшие масштабы ожидаемого природного явления, но в реальности оно имеет непрогнозируемую силу и влечет за собой необратимые последствия. Или вариантом такого события может стать наводнение, вызванное прорывом неотрмонтированной дамбы, произошедшим по вине человека.

На материке, где в некоторый временной промежуток прогнозируется природное явление, имеется локальная социальная сеть для общения и обмена информацией, предназначенная конкретно для жителей этой местности. Цель созданной социальной сети заключается в том, чтобы жители, взаимодействуя друг с другом, получали сведения о развивающейся природной ситуации и обстановке вокруг них. Общение в этой социальной сети не обладает приватностью – это аналог открытого чата или форума, где любой житель в любой момент времени может написать или ответить на сообщение любого пользователя. Если извлечь из данной социальной сети сообщения, то получится большая база данных в виде таблицы, структурированных по рассмотренным ранее атрибутам. Проведя анализ этих сообщений, можно прийти к выводам, ключевым из которых становится определение региона, в котором происходит природное явление; для этого понадобится атрибут локации. Сведения помогут экстренным службам не только более оперативно прибыть на требуемое место, но и знать заранее вид последствий произошедшего события и, соответственно, вид помощи, которую необходимо оказать. К тому же, есть возможность вычислить, кому именно нужна помощь, так как имеются данные об аккаунте пользователя. В таком случае люди смогут помочь своим родным и знакомым при первой возможности, или, находясь неподалеку от эпицентра, обезопасить себя, обладая информацией о происходящем вокруг.

Для подобного анализа потребуется поиск интересных нас связей в сообщениях. Это могут быть слова и словосочетания, которые связаны с конкретным природным явлением, или особенности ввода, например повторяющиеся символы в тексте сообщений. Фильтруя данные, обращая внимание на эти особенности, можно будет отделить относящиеся к делу сообщения от нерелевантных, представляющих собой общение пользователей на отвлеченные темы, неуместные объявления и спам. Для обработки вручную даже структурированного массива данных потребуются большие усилия и время, поэтому лучшим решением будет написание прикладной программы, выполняющей необходимый функционал.

Итоговые данные – набор из чисел, соответствующих определенным регионам с данными о пользователях, нуждающихся в помощи по категориям. Для их визуализации выбираются более удобные форматы для анализа.

Для оптимизации процесса за одну операцию выполняются два шага – отбор (фильтрация) сообщений по ключевым словам и их подсчет (сортировка, распределение по регионам).

В качестве входных параметров основной функции выступают:

1. Правило отбора: из первоначального массива данных фильтруются необходимые сообщения.

2. Правило подсчета: сообщения распределяются по категориям помощи и приписываются району, из которого были отправлены.

3. Выгрузка данных в файл. Выходные данные основной функции – множество строк формата:

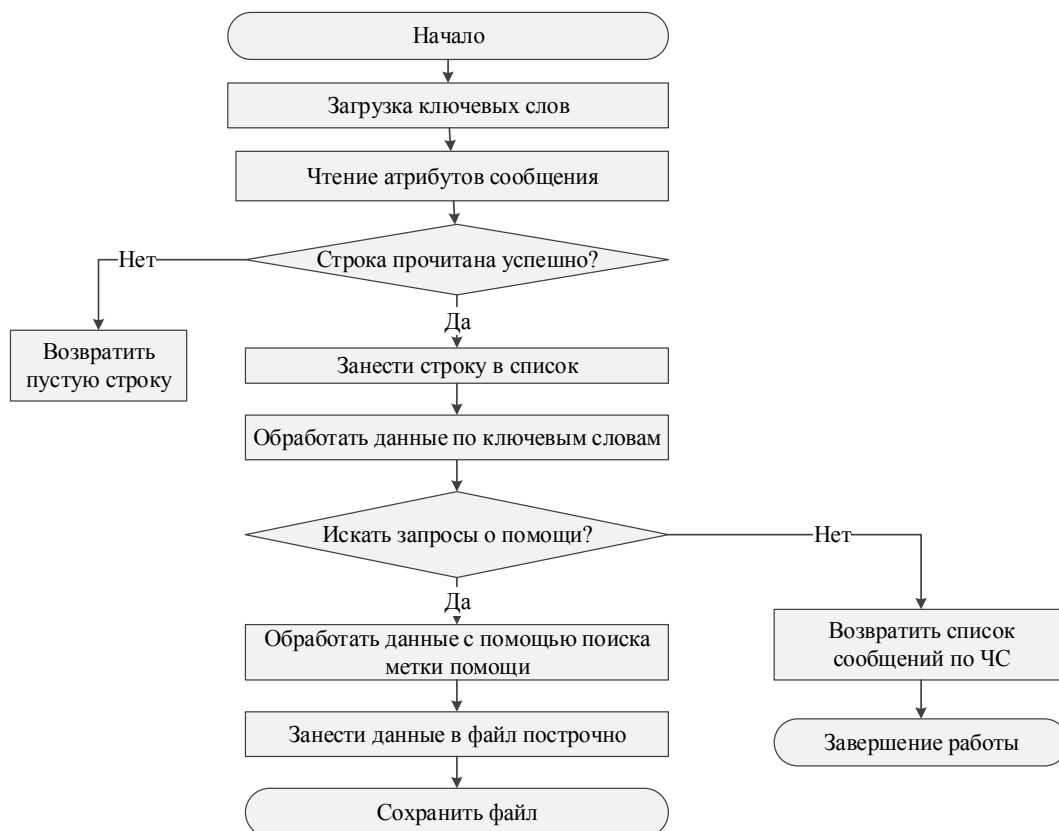


Рис. 1. Алгоритм поиска сообщений по ключевым словам и меткам

Fig. 1. Algorithm for searching messages by keywords and tags

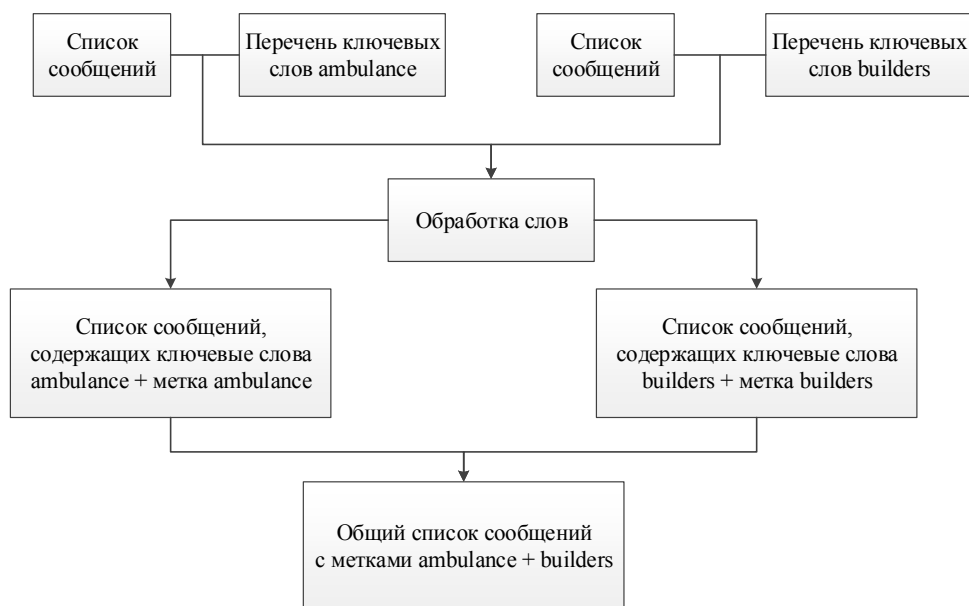


Рис. 2. Процесс обработки данных

Fig. 2. Data processing

«строка; число; число; ...; число», представляющих собой соответствие названия локации и числа пользователей, нуждающихся в конкретной помощи. Символ «;» разделяет категории помощи. Данные выгружаются по пути файла, заданного пользователем. Формат по умолчанию – csv.

На рис. 1 представлена принципиальная схема описанного алгоритма.

Исходными данными для анализа служит набор сообщений, написанных в период с 6.04.2019 по 10.04.2019, извлеченных из социальной сети города. В данный период времени в городе случилось землетрясение, которое вызвало реакцию людей в социальной сети. Данные представлены в формате .csv, где каждое сообщение обладает следующими характеристиками: дата и время отправки сообщения; район, из которого поступило сообщение; автор сообщения; текст сообщения. Число сообщений, извлеченных из социальной сети, составляет 41 868.

Для исследования был определен список ключевых слов, связанных с катастрофой. По этим словам происходит процесс фильтрации сообщений по релевантности, т. е. сообщения, содержащие в себе ключевые слова, отделяются от остальных, которые отменяются. К списку таких слов относятся: землетрясение, тряска, дрожание, колебание. Также существуют списки ключевых слов для каждого вида помощи: например, для медицинской помощи – ранение, перелом, кровь, ушиб и т. д.; для бригады устранения обвалов – разлом, обвал, завал и т. д.

Процесс обработки данных визуализирован на рис. 2.

Помимо этого, было добавлено условие проверки повторяющихся букв в сообщениях: при наличии в тексте более 3 одинаковых букв подряд сообщения также добавляются в категорию связанных с землетрясением, ведь при дрожи земли человек может несколько раз напечатать одни и те же символы и отправить сообщение, не задумываясь о корректности написания.

Город St. Nimmark делится на 19 районов, и каждый пользователь сети имеет возможность отправки сообщения из каждого из них.

Результатом работы программы должен служить файл с информацией о виде помощи и количестве пользователей, нуждающихся в помощи, по регионам (рис. 3). На первичном этапе программа отбирает сообщения, содержащие ключевые слова; затем распределяет среди этих сообщений категории помощи, после чего считает количество нуждающихся в помощи по категори-

ям помощи для каждого региона. Данные выгружаются в файл на компьютере. «Запустим» программу и проследим за каждым этапом.

Запуск реализуется функцией resources (). Данные на этом этапе выгружаются в файл. При входе в каталог обнаружен файл с данными, внутри – таблица, где 1-й столбец – это регион, 2-й – количество пользователей, нуждающихся в медицинской помощи, 3-й – количество пользователей, нуждающихся в помощи бригад разбора завалов (рис. 4).

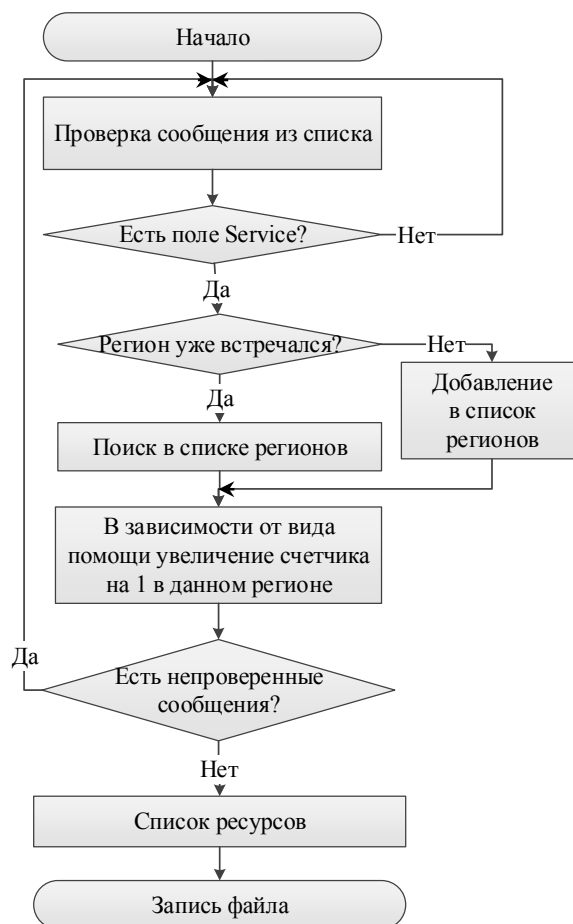


Рис. 3. Определение типа требуемой помощи и количества нуждающихся в ней пользователей в регионе

Fig. 3. Determination of the type of assistance required and the number of users in need in the region

West Part	3	4
Safe Town	5	6
Broadview	7	5
Oak Willow	4	4
Cheddarfo	7	5
Southton	7	8
Scenic Vist	4	6
Old Town	5	13

Рис. 4. Пример записанного файла (скриншот)

Fig. 4. Example of recorded file (screenshot)

Обсуждение результатов. По результатам тестирования на данных, представляющих собой набор сообщений жителей в социальной сети конкретного города, можно сделать вывод о работоспособности разработанного алгоритма. Однако для повышения его эффективности необходимо более точное определение ключевых слов и фраз для поиска сообщений, относящихся к ситуации. К примеру, шутки пользователей могут относиться к тематике и содержать ключевые слова. Эта информация будет воспринята напрямую, без учета правдивости опубликованной информации.

Выводы и заключение. Результаты предложенного подхода и его реализации могут быть использованы в различных социальных сферах, в частности в жизнеобеспечении, дорожной безопасности, противодействии террористическим актам. Дальнейшее развитие может быть направлено на повышение эффективности распределения ресурсов и производительности алгоритма посредством увеличения точности лексического анализа сообщений и параллельной обработки данных.

Финансирование. Работа выполнена при поддержке Гранта Президента Российской Федерации МК-918.2022.1.6.

Список литературы

1. Sharma R., Ahuja V., Alavi S. The future scope of netnography and social network analysis in the field of marketing // J. of internet commerce. 2018. Vol. 17, № 1. С. 26–45.
2. Zandian Z. K., Keyvanpour M. R. Feature extraction method based on social network analysis // Appl. Artificial Intelligence. 2019. Vol. 33, № 8. С. 669–688.
3. Political, economic, social, technological, legal and environmental dimensions of electric vehicle adoption in the United States: A social-media interaction analysis / R. Debnath, R. Bardhan, D. M. Reiner, J. R. Miller // Renewable and Sustainable Energy Reviews. 2021. Vol. 152. С. 111707.
4. Valuing sports services in urban parks: A new model based on social network data / P. Dai, S. Zhang, H. Hou, Y. Yang, R. Liu // Ecosystem Services. 2019. Vol. 36. P. 100891.
5. Valeri M., Baggio R. Social network analysis: Organizational implications in tourism management // International J. of Organizational Analysis. 2020. Vol. 29, № 2. P. 342–353.
6. Vikatos P., Gryllos P., Makris C. Marketing campaign targeting using bridge extraction in multiplex social network // Artificial Intelligence Rev. 2020. Vol. 53, № 1. P. 703–724.
7. Understanding the psycho-sociological facets of homophily in social network communities / R. S. Solomon, P. Y. K. L. Srinivas, A. Das, B. Gamback, T. Chakrabort // IEEE Computational Intelligence Magazine. 2019. Vol. 14, № 2. P. 28–40.
8. Detection of extremist ideation on social media using machine learning techniques / S. Mussiraliyeva, M. Bolatbek, B. Omarov, K. Bagitova // Intern. Conf. on Computational Collective Intelligence. Springer, Cham, 2020. P. 743–752.
9. Tsarenko O., Bekeneva Y., Novikova E. Analysis of the Messages from Social Network for Emergency Cases Detection // Congress on Intelligent Systems. Springer, Singapore, 2020. P. 35–47.

Информация об авторах

Бекенева Яна Андреевна – канд. техн. наук, доцент кафедры Вычислительной техники СПбГЭТУ «ЛЭТИ».

E-mail: yabekeneva@etu.ru

<https://orcid.org/0000-0002-7110-6000>

Байдуров Илья Александрович – студент СПбГЭТУ «ЛЭТИ».

E-mail: bayduroff@yandex.ru

References

1. Sharma R., Ahuja V., Alavi S. The future scope of netnography and social network analysis in the field of marketing // J. of internet commerce. 2018. Vol. 17, № 1. С. 26–45.
2. Zandian Z. K., Keyvanpour M. R. Feature extraction method based on social network analysis // Appl. Artificial Intelligence. 2019. Vol. 33, № 8. С. 669–688.
3. Political, economic, social, technological, legal and environmental dimensions of electric vehicle adoption in the United States: A social-media interaction analysis / R. Debnath, R. Bardhan, D. M. Reiner, J. R. Miller // Renewable and Sustainable Energy Reviews. 2021. Vol. 152. С. 111707.
4. Valuing sports services in urban parks: A new model based on social network data / P. Dai, S. Zhang, H. Hou, Y. Yang, R. Liu // Ecosystem Services. 2019. Vol. 36. P. 100891.
5. Valeri M., Baggio R. Social network analysis: Organizational implications in tourism management // International J. of Organizational Analysis. 2020. Vol. 29, № 2. P. 342–353.

6. Vikatos P., Gryllos P., Makris C. Marketing campaign targeting using bridge extraction in multiplex social network // Artificial Intelligence Rev. 2020. Vol. 53, № 1. P. 703–724.

7. Understanding the psycho-sociological facets of homophily in social network communities / R. S. Solomon, P. Y. K. L. Srinivas, A. Das, B. Gamback, T. Chakrabort // IEEE Computational Intelligence Magazine. 2019. Vol. 14, № 2. P. 28–40.

8. Detection of extremist ideation on social media using machine learning techniques / S. Mussiraliyeva, M. Bolatbek, B. Omarov, K. Bagitova // Intern. Conf. on Computational Collective Intelligence. Springer, Cham, 2020. P. 743–752.

9. Tsarenko O., Bekeneva Y., Novikova E. Analysis of the Messages from Social Network for Emergency Cases Detection // Congress on Intelligent Systems. Springer, Singapore, 2020. P. 35–47.

Information about the authors

Yana A. Bekeneva – Cand. Sci. (Eng.), Assistant Professor of the Department of Computer Technology, Saint Petersburg Electrotechnical University.

E-mail: yabekeneva@etu.ru

<https://orcid.org/0000-0002-7110-6000>

Ilya A. Baydurov – student, Saint Petersburg Electrotechnical University.

E-mail: bayduroff@yandex.ru

Статья поступила в редакцию 22.08.2022; принята к публикации после рецензирования 25.08.2022; опубликована онлайн 21.10.2022.

Submitted 22.08.2022; accepted 25.08.2022; published online 21.10.2022.
