Анализ открытых данных социальной сети с целью идентификации девиантных сообществ

Р.М. Михерский[1], Д.А. Кузнецов[1]

[1]Крымский федеральный университет имени В.И. Вернадского, пр. академика Вернадского 4, Симферополь, Россия, 295007

Аннотация. Разработана и программно реализована система анализа открытых данных социальной сети Вконтакте. Предложено два способа идентификации девиантных сообществ. Первый способ, по числу подписчиков сообщества, заблокированных социальной сетью за нарушение правил. Второй способ, по наличию общих подписчиков между исследуемым сообществом, и сообществом о котором точно известно, что оно девиантое. Экспериментально установлено, что второй способ идентификации девиантных сообществ дает лучший результат.

1. Постановка задачи

Анализ открытых данных социальных сетей является значимым направлением в области обработки больших данных. В частности, важной задачей, как для правоохранительных органов, так и для администраторов социальных сетей является выявление сообществ этих сетей, распространяющих общественно опасный контент. Обсуждению данной проблемы посвящено достаточно много работ публикуемых в последнее время. Работа [1] посвящена разработке метода оценки степени связанности профилей пользователей социальных сетей на основе открытых данных. Под степенью связанности профилей пользователей понимается вероятность знакомства владельцев профилей в реальной жизни. В работе [2] произведен обзор методов, которые обнаруживают демографические атрибуты пользователя из их профиля и сообщений. В работах [3,4] подробно рассмотрены формы девиантного поведения пользователей русскоязычного сегмента сети Интернет. В частности в работе [4] показано, что основной причиной девиантного поведения в социальных сетях является виртуальность и ананимность. В работе [5] по данным зарубежных источников проведен обзор основных методов анализа социальных сетей применительно к задаче выявления подозрительных и преступных сообществ.

К сожалению, чаще всего, выявление девиантных сообществ проводится в ручном режиме, зачастую, лишь по жалобам пользователей.

Целью данной работы явилась разработка методики выявления девиантых сообществ в социальной сети Вконтакте в автоматическом режиме. Для достижения этой цели было предложено два варианта поиска подобных сообществ.

2. Результаты

В первом варианте предложен и программно реализован следующий алгоритм поиска подобных сообществ. Для исследуемого сообщества определяется число подписчиков l

заблокированных социальной сетью за нарушение правил, а так же общее число подписчиков L данного сообщества. Находится коэффициент $k = \frac{l}{L}$. Предполагается, что если коэффициент k больше некоторого критического значения k_d, то исследуемое сообщество относится к девиантным.

Программная реализация представленного выше алгоритма была осуществлена на языке программирования Python. В ходе выполнения этой программы было случайным образом отобрано 50704 сообщества социальной сети Вконтакте. Из этих сообществ были отобраны те, численность подписчиков в которых 100 и более человек. Для каждого из данных сообществ был подсчитан коэффициент k. Далее все сообщества сортировались по величине этого коэффициента по убыванию. В таблице 1 представлены первые 20 сообществ из полученного списка.

Таблица 1. Сообщества с большим процентом заблокированных подписчиков.

№	Идентификационный номер сообщества	Число подписчиков в сообществе	Число заблокированных подписчиков	Процент заблокированных подписчиков от общего числа подписчиков сообщества, k•100%
1	172017411	104	101	97,1154
2	171896750	122	114	93,4426
3	41398959	107	98	91,5888
4	125043269	1017	904	88,8889
5	19613748	960	852	88,75
6	176328754	226	193	85,3982
7	148023353	495	419	84,6465
8	188941498	530	438	82,6415
9	23811356	1116	921	82,5269
10	164252296	152	123	80,9211
11	150230769	198	157	79,2929
12	130381011	200	157	78,5
13	154988787	410	317	77,3171
14	155397881	847	654	77,2137
15	149830913	107	81	75,7009
16	170030633	577	428	74,1768
17	***	174	129	74,1379
18	164288533	153	113	73,8562
19	143657800	424	312	73,5849
20	157513161	420	309	73,5714

С целью недопущения пропаганды девиантных сообществ, в этой таблице и далее в таблице 2, идентификационный номер всех таких сообществ заменяется символами «***».

Как видно из данного списка, в нем присутствует всего одно девиантное сообщество (сообщество под № 17). Это сообщество было отнесено к девиантным в связи с присутствием в нем материалов порнографического характера.

Таким образом, гипотеза о том, что в девиантых сообществах процент заблокированных пользователей больше, чем в не девиантых не нашла экспериментального подтверждения.

Второй вариант поиска девиантых сообществ основан на следующем алгоритме: Находится одно сообщество, для которого точно известно, что оно девиантное. Для данного сообщества определяется список подписчиков. У каждого из этих подписчиков определяются те сообщества, на которые он подписан. Для каждого из сообществ этого списка определяется количество подписчиков являющихся так же подписчиками исследуемого девиантого сообщества. Предполагается, что достаточно большое число сообществ из данного списка тоже

будут девиантными. Данный алгоритм был программно реализован с помощью языка программирования Python.

Для проверки работоспособности, этой программы было выбрано девиантое сообщество «Мама Анархия» с идентификационным номером 177615404. Данное сообщество занимается популяризацией идей анархизма и имеет 32097 подписчика. Время обработки данных составило – 18 часов. Подписчики этого сообщества подписаны так же на 940512 других сообществ. Все они были отсортированы в порядке убывания по числу пользователей, которые также подписаны и на сообщество «Мама анархия». В таблице 2 представлено первые 20 сообществ из этого списка.

Таблица 2. Сообщества, подписчики которых являются также подписчиками сообщества «Мама анархия».

№	Идентификационный номер сообщества	Число подписчиков в сообществе	Число подписчиков, которые являются также подписчиками сообщества «Мама Анархия»
1	***	5539982	15035
2	91050183	9356399	12924
3	***	707327	12712
4	159146575	1162785	12521
5	***	563784	11987
6	***	4403183	11644
7	***	2768306	11317
8	***	2508543	11246
9	57846937	11275065	11224*
10	***	2684988	11154
11	***	2586853	11145
12	150550417	937052	10916
13	149094324	2076903	10832*
14	30316056	1809325	10451
15	66678575	4976245	10299
16	12353330	3555825	10167
17	154168174	1264550	10145
18	173556111	641480	10005
19	***	3802683	9576
20	133180305	3116645	9540

Как видно из этой таблицы, из 20 сообществ, представленного списка, 9 относятся к девиантным. Основные причины того, что эти сообщества отнесены к девиантым: пропаганда насилия, критика существующего конституционного строя, использование ненормативной лексики.

3. Заключение

Таким образом, второй способ идентификации девиантных сообществ является гораздо более эффективным, чем первый. Эта методика выявления девиантых сообществ в автоматическом режиме может быть применена не только в социальной сети Вконтакте но и в других социальных сетях. Отметим также, что второй метод может быть применен не только для поиска девиантных сообществ, но и при поиске сообществ связанных с изучаемым сообществом, например, при маркетинговых исследованиях.

4. Литература

[1] Катаева, В.А. Методы оценки степени связанности профилей пользователей социальной сети на основе открытых данных / В.А. Катаева, И.С. Пантюхин, И.В. Юрин // Открытое образование. – 2017. – Т. 21, № 6. – С. 14-22.

[2] Гомзин, А.Г. Методы построения социально-демографических профилей пользователей сети Интернет / А.Г. Гомзин, С. Д. Кузнецов // Труды ИСП РАН. – 2015. – Т. 27, № 4. – С. 129-142.
[3] Бакланцева, А.А. Трансформация социальных норм и девиаций в русскоязычной сети Интернет // Изв. вузов. Сев. Кавк. регион. Обществ. науки. – 2014. – № 3. – С. 21-25.
[4] Черенков, Д.А. Девиантное поведение в социальных сетях: причины, формы, следствие // Nauka-Rastudent.ru. – 2015. – № 7(19). – С. 29.
[5] Басараб, М.А. Обнаружение противоправной деятельности в киберпространстве на основе анализа социальных сетей: алгоритмы, методы и средства (обзор) / М.А. Басараб, И.П. Иванов, А.В. Колесников, В.А. Матвеев // Вопросы кибербезопасности. – 2016. – № 4(17). – С. 11-19.

Analysis of open data of a social network in order to identify deviant communities

R.M. Mikherskii[1], D.A. Kuznetsov[1]

[1]V.I. Vernadsky Crimean Federal University, Vernadskogo Prospekt 4, Simferopol, Russia, 295007

Abstract. The system of analysis of open data of the social network Vkontakte is developed and programmatically implemented. Two ways of identification of deviant communities are proposed. The first way is by the number of community subscribers blocked by the social network for violating the rules. The second way, by the presence of common subscribers between the studied community, and the community about which it is precisely known that it is deviant. It is experimentally established that the second method of identification of deviant communities gives the best result.