Анализ контента социальной сети на примере квестовой игры суицидального характера, направленной на детей и подростков

А.И. Петров, О.С. Смирнова, Б.Б. Чумак

Аннотация **– В данной статье рассматривается разработка программного модуля для анализа открытых данных социальных сетей с целью выявления первоисточника заданного контента на примере квестовой игры суицидального характера, направленной на детей и подростков, в социальной сети «ВКонтакте». Описаны исходные данные и их анализ в рамках поставленной задачи, рассмотрен процесс наполнения базы данных, а также проблемы, возникшие во время разработки модуля. В статье представлены результаты анализа данных по рассматриваемому примеру.**

Ключевые слова **– социальные сети, интеллектуальный анализ данных, разработка программного модуля.**

I. Введение

Одна из актуальных на сегодняшний день тем для анализа данных является анализ открытых данных социальных сетей. Цели такого анализа: выявление разного рода неявных закономерностей в данных, что может поспособствовать в описании поведений как отдельных пользователей, так и целых сообществ.

Примером целого набора задач для анализа данных социальных сетей является квестовая игра суицидального характера «Синий кит», пик популярности которой пришелся на конец 2016 – начало 2017 года в социальной сети «ВКонтакте». Подростки делали записи в социальной сети, благодаря которым так называемые «кураторы» указанной игры связывались с ними и последовательно предлагали ряд заданий, последним из которых было совершение суицида.

С целью выявления пользователей, попадающих под влияние рассматриваемой игры, а также пользователей, способствующих распространению «опасного» контента, был спроектирован и разработан программный модуль для анализа открытых данных пользователей социальных сетей.

Статья получена 30.05.2017 г.
Исследование выполнено федеральным государственным бюджетным образовательным учреждением высшего образования «Московский технологический университет» (МИРЭА) за счет гранта Российского фонда фундаментальных исследований (проект №16-37-00492).
А.И. Петров, МИРЭА (e-mail: au.lewka@gmail.com)
О.С. Смирнова, МИРЭА (e-mail: mail.olga.smirnova@yandex.ru)
Б.Б. Чумак, к.т.н., МИРЭА (e-mail: chumak@mirea.ru)

Задачи данного модуля составляют:

– поиск первоисточников заданного контента;

– определение общего контента профилей-первоисточников.

Следует отметить, что под первоисточником контента подразумевается пользовательский профиль, на котором впервые был опубликован анализируемый контент.

II. Исходные данные

Исходными данными для решения данной задачи являются открытые данные социальных сетей и набор хеш-тегов: #синийкит, #4:20, #тихийдом, которые используются пользователями социальной сети «ВКонтакте» как показатель желания участия в «игре».

Очевидно, что все данные социальной сети использовать не получится из-за их колоссального объема, поэтому перед началом сбора необходимо установить некоторые ограничения. Например, можно ограничить группу пользователей, данные которых предполагается использовать в анализе, демографическими показателями: возраст, пол, страна, город. Таким образом, было решено выделить пользователей от 10 до 18 лет, проживающих в Москве, что позволило ограничить количество рассматриваемых профилей с нескольких сотен миллионов до 2.5 миллионов.

Для данной конкретной задачи социальная сеть является единственным источником данных, но другие задачи могут потребовать и дополнительные источники.

III. Разработка

Как правило работа по анализу данных начинается с поиска ключевых атрибутов (feature engineering). Суть процесса заключается в том, что из всего имеющегося набора атрибутов происходит выбор только тех, которых действительно могут быть полезны при анализе.

Для решения поставленной задачи были исследованы атрибуты моделей данных для профиля и для записи. Часть атрибутов было решено не использовать вовсе, например, ссылки на фотографии

профиля, поэтому такие данные не будут выгружаться из базы данных (БД) социальной сети в рамках решения поставленных задач. Обработку части атрибутов моделей было решено проводить после выгрузки из БД социальной сети, перед загрузкой данных в локальную БД. Например, текстовые поля «Любимые цитаты», «Любимые фильмы» в рамках данной задачи обрабатываются таким образом, что в локальную БД записывается только факт наличия или отсутствия этих данных. Это позволит сократить объем данных, а также упростить анализ профилей.

Следующим этапом работы являлся сбор данных, включающий разработку и запуск скриптов для выгрузки данных из социальной сети с использованием специального программного интерфейса, а также загрузку этих данных в локальную базу данных.

В процессе работы было загружено 2.2 тысяч записей с анализируемыми хэш-тегами и аналогичное количество профилей пользователей. Первая запись датируется 2013-09-10, последняя – 2017-05-05. График динамики количества публикаций анализируемого контента по месяцам приведен на рисунке 1.

Рисунок 1 – Динамика публикаций

Количество записей по полу пользователя распределилось следующим образом:

– 52% пользователей – женского пола;

– 23% пользователей – мужского пола;

– 25% пользователей – не указали свой пол.

Соответствующая диаграмма представлена на рисунке 2.

Описанные этапы разработки контекстно-независимые, то есть они выполняются при решении любых задач по анализу данных с учетом лишь специфики интересующих данных и источника информации. Следующие же этапы, в общем случае представляющие собой непосредственный анализ данных, зависят в большей степени от поставленной задачи.

Рисунок 2 – Количество записей по полу пользователя

В данном случае, для решения задачи поиска первоисточника контента, следующим этапом является выбор некоторой части найденных записей, опубликованных раньше других. При этом, если одна запись является дочерней к другой записи («репостом»), то берется родительская, исходная запись. Таким образом, авторов выбранных записей будем считать первоисточником искомого контента.

Для выборки было решено взять 20% от всего количества загруженных записей, что составило 556 записей. Самая ранняя запись с хеш-тегом «#забери меня» была сделана 10 сентября 2013 года.

Однако после поверхностного анализа выбранных записей было решено ограничить дату первой записи выборки первым ноябрём 2015 года – датой, которая считается ориентировочным началом истории «групп смерти».

Последующий поверхностный анализ отобранных записей показал, что 80 – 90% записей не причастны к игре «Синий кит», а являются обыкновенными записями с указанием хеш-тегов, попавшими в список «опасных». По этой причине, из всех отобранных записей были выбраны только те, которые действительно имеют отношение к тематике игры. В итоговой выборке оказалось всего 29 записей, данного количества для полноценного анализа общего контента пользовательских профилей, которые можно считать первоисточниками записей, слишком мало. Тем не менее, такой функционал был реализован в программном модуле анализа данных, и он показал следующие результаты:

– в 97% профилей источников не была указана дата рождения;

– в 90% профилей источников не было указано отношение к алкоголю и курению;

– в 75% профилей количество таких параметров как «подписчики», аудиозаписи, фотографии, видео, друзья, записи, видео с пользователями, группы равно 0, при этом профили не являются скрытыми и деактивированными, у них открыта возможность комментирования записей на «стене», установлена фотография пользователя и указано короткое имя страницы пользователя.

По полученной итоговой выборке можно сделать вывод о том, что на текущий момент в социальной сети содержится очень мало записей с хеш-тегами #синийкит, #4:20, #тихийдом, свойственными квестовой игре суицидального характера «Синий кит»: условно 70% всех записей не имеют отношения к игре, 25% – являются записями, призывающими не участвовать в игре и остальные 5% – записи, соответствующие тематике игры. Следует отметить, что при этом многие СМИ утверждали о достаточно большой волне популярности данной игры среди детей и подростков. Из этого можно сделать вывод, что либо администрация социальной сети «ВКонтакте» активно удаляет записи с «опасными» хэш-тегами, по которым проводился анализ, либо в действительности информация об активном вовлечении детей и подростков в упомянутую игру преувеличена в СМИ. Но чтобы делать окончательный вывод, необходимо проведение дополнительных исследований: как и по другим наборам хэш-тегов, так и на основе более подробного анализа профилей и активностей пользователей: размещаемого текстового и аудио контента, наличию соответствующих игре информационных образов на изображениях, временной активности и др.

IV. Проблемы, возникшие в процессе разработки

Первой проблемой стал большой объем текстовых данных в профилях пользователей. Решением была замена текстовых данных профилей на флаги истинности, указывающие на наличие или отсутствие тех или иных данных, что позволило до минимума сократить объемы текстовых данных профилей.

Другой проблемой стали ограничения API на получение данных из социальной сети «ВКонтакте». Решением этой проблемы было деление на части запросов на получение больших массивов данных, таким образом время получения было увеличено, но при этом все ограничения удалось преодолеть.

Наконец последней проблемой стало большое количество лишних записей, не относящихся к заданной тематике. Для решения были выставлены ограничения по времени размещения анализируемых записей, после чего отобранные записи были просмотрены на соответствие и связь с заданной темой исследования.

V. Заключение

В данной статье описана разработка модуля для анализа открытых данных пользователей социальных сетей, с помощью которого было проведено исследование контента социальной сети «ВКонтакте». В качестве примера для анализа был использован контент, присущий квестовой игре суицидального характера «Синий кит». Целью анализа являлось выявление профилей-первоисточников заданного контента и анализ общего контента выявленных профилей. В работе описаны полученные результаты анализа. Следует отметить, что текущая версия разработанного модуля предполагает работу в полуавтоматическом режиме, что в дальнейшем планируется усовершенствовать, а также планируется разработать графический интерфейс для работы с модулем и программный интерфейс, что позволит интегрировать данный модуль в разрабатываемую систему анализа [3, 4, 5, 6] и любую другую систему.

VI. Список используемых источников

1. Синий кит (игра) – Википедия [электронный ресурс] // URL: https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D0%BD%D0%B8%D0%B9_%D0%BA%D0%B8%D1%82_(%D0%B8%D0%B3%D1%80%D0%B0) (дата обращения: 30.04.2017)

2. Документация | Разработчикам [электронный ресурс] // URL: https://vk.com/dev/manuals (дата обращения: 2.04.2017)

3. Смирнова О.С. 1, Петров А.И. 2, Бабийчук Г.А. Основные методы анализа, используемые при исследовании социальных сетей // Современные информационные технологии и ИТ-образование. Т.12 (№ 3), часть 1, 2016, с. 151 – 158.

4. В.В. Баранюк, А.Д. Десяткова, О.С. Смирнова. Подходы к определению психоэмоциональных особенностей информационного образа пользователя социальных сетей // International Journal of Open Information Technologies. Том 4, № 8 (2016), с. 61 – 65.

5. А.С. Алымов, В.В. Баранюк, О.С. Смирнова. Детектирование бот-программ, имитирующих поведение людей в социальной сети «Вконтакте» // International Journal of Open Information Technologies. Том 4, № 8 (2016), с. 55 – 60.

6. О.С. Смирнова, В.В. Шишков. Выбор топологии нейронных сетей и их применение для классификации коротких текстов // International Journal of Open Information Technologies. Том 4, № 8 (2016), с. 50 – 54.

An analysis of the content of a social network as exemplified in a suicide themed quest game, aimed at the children and adolescents

A.I. Petrov, O.S. Smirnova, B.B. Chumak

***Abstract* – This article deals with software module development for analytics of social network data with aim to detecting of original source of content with example of quest game of a suicidal nature aimed at children and adolescents in social network «VK». The article describes the initial data and the analysis of that data, the article deals with the process of the database filling and with the problems that occurred during the development of the software module. Also the article presents the results of data analysis for the task.**

***Keywords* – social networks, data mining, software development.**