УДК 004.89

ИССЛЕДОВАНИЕ АЛГОРИТМОВ АНАЛИЗА ИНФОРМАЦИИ В СОЦИАЛЬНЫХ СЕТЯХ

Ищукова Е.А., Салманов В.Д., Калябин А.А., Крюк Д.Н.

Южный федеральный университет, Институт компьютерных технологий и информационной безопасности, Таганрог, e-mail: uaishukova@sfedu.ru

С развитием информационных технологий и расширением информационного пространства сети Интернет контроль качества контента становится все менее возможным. В социальных сетях распространяется огромное количество информации: от личных фото и смешных изображений до рекламы, призывов к суициду и т.д. Распространяется явление кибербуллинга, личных оскорблений. В связи с этим актуальной становится задача анализа информации различными автоматизированными средствами. В данной работе авторы описывают опыт разработки и использования различных технологий анализа данных, расположенных в открытом доступе в социальных сетях. При подготовке к исследованию был разработан парсер для сбора и обработки первичных данных из сети. В ходе исследования были проанализированы различные подходы к построению нейронных сетей для анализа контента. Приводятся результаты работы алгоритма по построению психоэмоционального портрета пользователя на основе теста «Большая пятерка», с помощью нейронных сетей удалось установить некоторое соответствие численных параметров пользователя социальной сети с его психологическими характеристиками. Описаны результаты изучения динамики эмоционального состояния пользователей путем анализа позитивного и негативного контента на странице аккаунта, приведены результаты использования нейронной сети на основе Наивного байесовского классификатора для анализа тональности текстов, размещаемых в социальных сетях.

Ключевые слова: социальные сети, семантический анализ, нейронные сети, негативный контент, психологический портрет

RESEARCH OF ALGORITHMS FOR ANALYZING INFORMATION IN SOCIAL NETWORKS

Ishchukova E.A., Salmanov V.D., Kalyabin A.E., Kryuk D.N.

Southern Federal University, Institute of Computer and Information Security, Taganrog, e-mail: uaishukova@sfedu.ru

With the development of information technologies and the expansion of the Internet information space, content quality control becomes less and less possible. In social networks, a huge amount of information is distributed from personal photos and funny images to advertising, calls for suicide, etc. the phenomenon of cyberbullying and personal insults is Spreading. In this regard, the task of analyzing information by various automated means becomes urgent. In this paper, the authors describe the experience of developing and using various data analysis technologies located in open access in social networks. In preparation for the study, a parser was developed for collecting and processing primary data from the network. The study analyzed various approaches to building neural networks for content analysis. The results of the algorithm for constructing a psychoemotional portrait of the user based on the «Big five» test are presented. using neural networks, it was possible to establish some correspondence between the numerical parameters of the social network user and his psychological characteristics. Describes the results of studying the dynamics of emotional States of users by analyzing the positive and negative content on the account page, the results of the use of a neural network-based Naive Bayes classifier for sentiment analysis of texts posted in social networks.

Keywords: social networks, semantic analysis, neural networks, negative content, psychological portrait

Современные темпы развития информационно-коммуникационных технологий настолько высоки, что оценить все их преимущества и недостатки своевременно не всегда становится возможным. Развитие социальных сетей влечет за собой создание огромного информационного поля. Причем объемы контента увеличиваются настолько быстро, что ручной мониторинг его качества становится невозможным. Становится актуальной задача автоматизации качественной оценки контента. Особую важность это проблема приобретает в контексте использования сети Интернет и социальных сетей детьми и подростками. Учащаются случаи кибербуллинга эмоционально не созревшего поколения, пропаганды насилия, экстремизма, суицида и прочих негативных неконтролируемых явлений в Сети. Наиболее популярными методами автоматической обработки больших массивов данных являются различные классификаторы, тем или иным образом использующие в своей основе нейронные сети.

Цель исследования: исследовать алгоритмы анализа данных в социальных сетях.

Материалы и методы исследования: сверточные нейронные сети, парсер, Наивный байесовский классификатор, ПЭВМ Intel(R) Core(TM) i5-7300HQ CPU 2.50GHz, язык программирования Python.

В ходе подготовки к реализации были выделены следующие ключевые алгоритмы, требующие реализации:

– программа для получения данных с открытых страниц пользователей социальной сети «ВКонтакте» и их дальнейшей разметки;

– программа для анализа открытых данных с целью выявления изменения его психоэмоционального состояния;

– программа для анализа открытых данных с целью выявления негативного контента на странице.

В настоящее время анализ данных из социальных сетей относят к анализу больших данных (BigData). Анализ больших данных затрудняется в первую очередь из-за того, что все данные разрозненны, имеют различную структуру и назначение. В настоящее время не существует универсальных алгоритмов, которые бы позволили бы проводить полный анализ профиля пользователя социальной сети. На ресурсе [1] собраны 35 наиболее известных автоматизированных средств, направленных на анализ данных социальных сетей. В большинстве своем эти средства оценивают контент количественно (сколько фото, видео, аудио, постов на странице пользователя), по времени активности пользователя, по наиболее часто употребляемым словам. Существуют исследования, направленные на выявление суицидально настроенных групп в социальной сети «ВКонтакте», но их работа основана на поиске наиболее распространенных хештегов [2]. Ни одно из представленных средств не проводит комплексный анализ, не отслеживает психоэмоциональное состояние человека, а значит, не может быть использовано для целей настоящего исследования.

Актуальной становится задача, которая заключается в разработке парсера. Парсер – это механизм, который позволяет извлекать данные из какого-либо источника, в нашем случае – социальной сети. Согласно цели проекта парсер в первую очередь технически разработан с учетом специфики социальной сети «Вконтакте», но его алгоритмы могут быть легко адаптированы для других платформ. В ходе разработки парсера были задействованы методы VK API, которые позволяют получать данные со страниц пользователей. Использование расширенных методов дало возможность проводить синтаксический анализ информации. Скрипт запрашивает профиль пользователя и возвращает подробную информацию о нем: имя, фамилию, возраст, дату рождения, семейное положение, имя и ссылку на партнёра, образование, работу, должность, записи на странице и прочее. Вся получаемая информация публична и берётся из открытых данных профиля, к закрытым данным скрипт не имеет доступа, что не совсем удобно, но не нарушает приватности пользователей.

Другие задачи были решены с использованием нейронных сетей. Нейросеть представляет собой модель, построенную по принципу биологической нейронной сети. Она состоит из системы взаимосвязанных процессоров, каждый из которых принимает сигнал от других процессоров, обрабатывает и передает его другим процессорам. Архитектура нейронной сети будет определять характер связи между процессорами (нейронами). Существует большое разнообразие видов нейронных сетей, сильно различающихся по сложности реализации, обучению, связи нейронов.

В ходе анализа данных, расположенных в социальной сети в открытом доступе, можно собрать информацию о психологических чертах пользователя. Под термином «психологический портрет» понимается совокупность в той или иной мере присущих человеку признаков: интроверсия/экстраверсия, уживчивость (доброжелательность), сознательность (добросовестность), нервозность/эмоциональная стабильность, открытость опыту (интеллект).

Именно эти характеристики оцениваются «Большой пятеркой» [3], это метод психологического анализа, в котором Маккрае и Коста предложили оценивать индивидуальные различия людей с учетом их биологических свойств. Этот метод вполне подходит для формирования психологического портрета, так как каждый из пяти факторов, описанных выше, является самостоятельной чертой характера. В дополнение к этому такой метод дает довольно низкую погрешность результатов и подходит для проведения масштабного анализа данных.

Следует учитывать специфику изучаемой социальной сети, для автоматического анализа будут использованы следующие параметры: размещаемые посты, количество изображений и фотографий на странице, аудио/видеозаписи, количество друзей и подписчиков, группы и сообщества, на которые подписан пользователь.

Для классификации наиболее подходящими оказались нейронные сети прямого распространения, одна из таких сетей – сеть радиально-базисных функций. Ее отличительным списком является применение монотонно возрастающих и монотонно убывающих с отдалением от центральной точки функций. Это позволит классифици-

ровать информацию на нужную и ненужную и производить основе этого анализ; кроме того, отличительным свойством сетей прямого распространения является обучение методом обратного распространения ошибки, когда на вход сети приходит большое количество входных и выходных данных, а ошибка заключается в разнице между входом и выходом [4]. В результаты мы получаем схему взаимодействия между входными и выходными данными. В табл. 1 показано, как каждый входной параметр влияет на каждый результирующий. Это достигнуто в результате сравнения весов, сформированных после обучения сети, представлено влияние каждого из входных параметров на каждый результирующий параметр.

Для построения нейронной сети по определению эмоционального состояния пользователя на сверточном уровне были использованы фильтры с высотой 2, 3, 4, 5 и созданы по 10 слоев для каждой высоты фильтра. Функцией активации является ReLU. Достаточно полно описан подбор фильтров в статье [5].

После работы сверточных слоев из карт признаков извлекалась наиболее значимая информация. Далее происходило соединение всех n-грамм в общий вектор признаков (слой объединения), который пересылается в следующий скрытый слой с 30 нейронами. В конце итоговая карта посылается на выходной слой с сигмоидальной функцией активации. Итог обучения нейронной сети представлен в табл. 2. Результаты работы программного комплекса представлены на рис. 1.

Для решения задачи определения тональности текста в целом и поиска негативного контента в частности были изучены существующие программные продукты, разработан собственный алгоритм оценки тональности текста, сделано сравнение их эффективности.

Один из вариантов оценки тональности текста – построение нейронной сети с помощью программной библиотеки TenzorFlow. При построении словаря каждому новому слову присваивался его уникальный индекс, тем самым получался массив длиною n. Затем входной текст разбивался на такие же слова и представал перед нейронной сетью в виде бинарного вектора длиной n, у которого на месте совпадающих слов была единица, в остальных позициях – нули.

Второй слой нейронной сети представлял собой 125 нейронов, третий – состоял из 25 нейронов, каждый из которых будет связан с каждым нейроном из входящего слоя. На выходе получались два значения, сумма которых сводилась к единице, таким образом, они характеризовали тональность текста в процентах.

Таблица 1

Влияние входных параметров на факторы «Большой пятерки»

	Интроверсия/ экстраверсия	Уживчивость, доброжела-тельность	Сознательность, добросовестность	Нервозность	Открытость опыту, интеллект
Размещаемые посты	0,12	0,05	–0,15	0,3	0,2
Количество изображений и фотографий	–0,23	–0,09	0,24	0,09	–0,18
Количество аудио/видеозаписей	0,06	–0,02	0,08	0,13	0,27
Количество друзей и подписчиков	0,1	–0,16	0,14	–0,3	0,04
Группы и сообщества	0,08	0,19	0,03	–0,01	0,01

Таблица 2

Итог обучения нейронной сети

Метка класса	Точность,%	Полнота,%	Наивысший показатель,%	Количество объектов
Negativ	83.194	83.243	83.218	22457
Positiv	84.089	84.040	84.064	22313
avg / total	83.142	83.142	83.142	44770

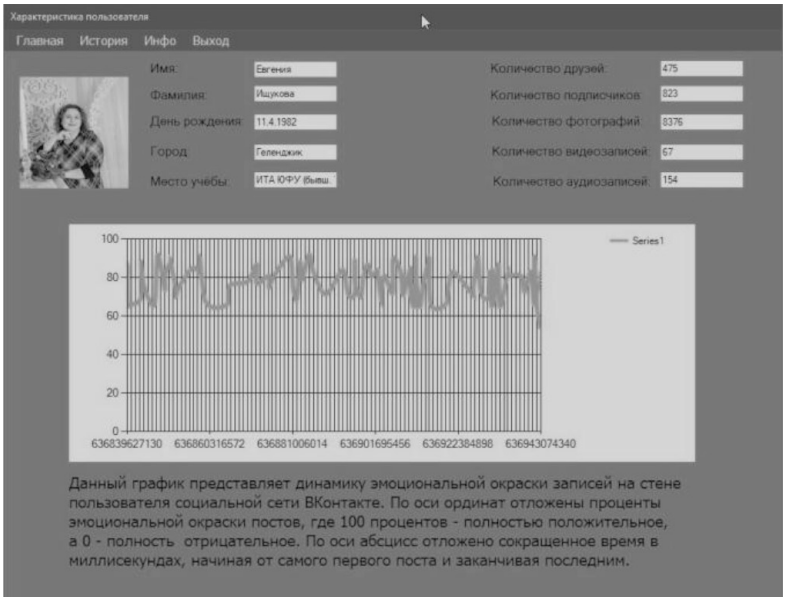

Рис. 1. Пример работы программы

Формирование словаря велось на базе размеченных позитивных и негативных постов, после чего словарь был сокращен до 5000 самых популярных слов. Это было сделано для того, чтобы отсеять слова, которые встречались редко, так как их тон при обучении будет определен однозначно. Обучение такой сети проводилось на выборке в 200 000 постов и на тестовой выборке давало точность порядка 94%. Это слишком высокий показатель для нейронной сети, так как она стала слишком умной и начала использовать самые простые правила. Поэтому, например, пост с большим количеством одиночно позитивных слов, но в целом негативный, считала позитивным.

Изучение решений для тонового анализа контента в социальной сети привело к проекту [6]. В основе проекта лежит Наивный байесовский классификатор. С помощью классификатора рассчитывается вероятность принадлежности к позитивному или негативному классу тональности, при этом допускается, что признаки в классе могут быть независимы. Это дало возможность сделать предположение о высоких показателях качества данного метода для анализа текстов в социальной сети.

Как и предыдущие алгоритмы, этот использует разделение текста на N-граммы (униграммы, биграммы и триграммы) для классификации текста. Дальнейшее использование формулы ΔTF–IDF позволяет выяснить, в скольких позитивных и негативных текстах встречается конкретный N-грамм. Разница этих значений будет характеризовать тональность этого N-грамма и, соответственно, всего текста.

Автором алгоритма приводится тестовая выборка, результаты работы над которой действительно говорят о наибольшей эффективности конкретного алгоритма. Описанный выше алгоритм оказался наиболее подходящим для решения задачи семантического анализа записей на страницах пользователей. Однако в ходе его использования на реальных задачах его показатели снизились. При работе с социальной сетью был разработан API, который считывал текст из последних ста постов на странице. Сформированные наборы текстов влияли на некорректное поведение алгоритма, что привело к необходимости его совершенствования. В ходе оптимизации решения были достигнуты показатели, близкие к заявленным (рис. 2).

В табл. 3 представлены некоторые результаты сравнения двух алгоритмов – разработанного самостоятельно и адаптированного существующего. Результаты получены путем оценивания тональности текста на специально созданных страницах в социальной сети (vk.com/id456817351, vk.com/id456827820).

В таблице приведены и выделены некоторые расхождения в результатах работы алгоритмов. Это связано с ошибками работы нейронных сетей, используемых классификаторов. Как и говорилось выше, сети не могут дать абсолютно идеальные показатели, и количество обнаруженных ошибок не выходит за рамки статистических показателей работы сетей.

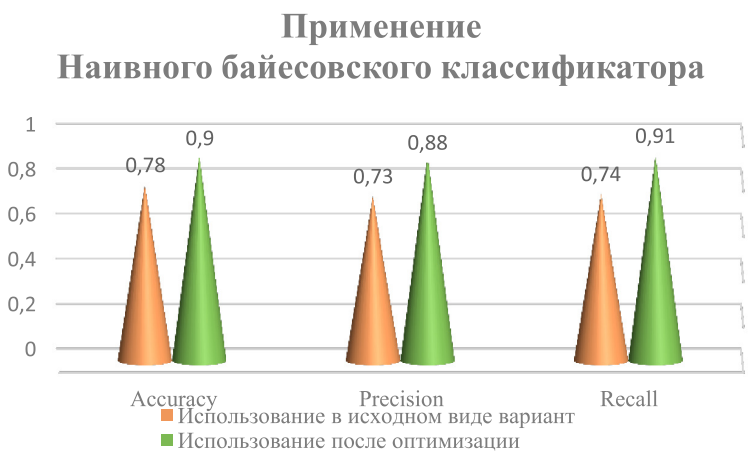

Рис. 2. Сравнение работы НБК.
Примечания: Accuracy (доля правильных ответов) = (P/N),
Precision (точность) = (TP/(TP + FP)) и Recall (полнота) = TP / (TP+FN).
P – позитивные; N – негативные; TP – верно определенные позитивные;
FP – ложно определенные позитивные; FN – ложно определенные негативные

Таблица 3

Сравнение результатов работы алгоритмов

Текст поста на странице	Разработанный алгоритм	Адаптированный алгоритм
Печаль, ошибка, я сдаюсь	0	0
Мы сами уничтожаем то, что создали	**1**	**0**
Ночь – время для дурацких мыслей	0	0
черный цвет всегда в моде	**0**	**1**
Ужасная погода, идет дождь!	0	0
С 14 марта переезжаю в новую квартиру! На этот раз остаюсь надолго...	1	1
любовь, улыбка, радость, счастье	1	1
улыбайтесь миру, и мир улыбнется вам в ответ	1	1
Дети – цветы жизни!	1	1
от улыбки станет всем светлей!	**1**	**0**

Результаты исследования и их обсуждение

В ходе работы были решены задачи обучения нейронных сетей. Теперь сети обладают информацией о соответствии количественных показателей, которые можно получить со страницы профиля, действительному психологическому портрету человека, тональности текста на странице его профиля. В ходе работы были сделаны следующие выводы:

– в случае успешного тестирования системы удавалось получить близкий к ожидаемому результат, при анализе дисперсии выходных значений был замечен либо достаточно близкий к математическому ожиданию результат, либо абсолютно неточный по нескольким характеристикам результат;

– некоторые возрастные категории отмечаются менее поддающимися анализу ввиду неполной или недостоверной информации в профиле социальной сети;

– уменьшение итогового количества слоев нейронной сети приводило к существенному уменьшению точности результата, дальнейшее увеличение количества слоев не приводило к значительному улучшению точности, но перегружало систему;

– существующие разработки требуют оптимизации под каждое конкретное программное решение для достижения наилучшего результата.

Заключение

В результате работы был разработан и реализован многомодульный программный продукт, который собирает, анализирует и представляет данные в удобном для конечного пользователя формате.

Работа выполнена при поддержке гранта РФФИ и ЭИСИ № 19-011-33006 «Разработка и исследование методов и подходов, направленных на формирование положительных культурно-нравственных качеств подростка посредством использования современной коммуникационной среды».

Список литературы

1. 35 инструментов для аналитики социальных сетей [Электронный ресурс]. URL: https://popsters.ru/blog/post/35-instrumentov-dlya-analitiki-socsetey (дата обращения: 06.03.2020).

2. Петров А.И., Смирнова О.С., Чумак Б.Б. Анализ контента социальной сети на примере квестовой игры суицидального характера, направленной на детей и подростков // International Journal of Open Information Technologies. 2017. Vol. 5. no 6. С. 16–18.

3. Большая пятерка (Big five). Пятифакторный личностный опросник (Р. МакКрае, П. Коста). Методика диагностики личностных факторов темперамента и характера (5PFQ). [Электронный ресурс] URL: https://psycabi.net/testy/388-test-bolshaya-pyaterka-pyatifaktornyj-lichnostnyj-oprosnik-r-makkrae-p-kosta-metodika-diagnostiki-lichnostnykh-faktorov-temperamenta-i-kharaktera-5pfq (дата обращения: 06.03.2020).

4. Liben-Nowell D., Kleinberg J. The link-prediction problem for social networks. J. Am. Soc. Inform. Sci. Technol. 58. P. 1019–1031. DOI: 10.1002/asi.20591, 2007.

5. Zhang Y., Wallace B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification arXiv:1510.03820. 2015.

6. Text tone analyzer [Electronic resource]. URL: https://github.com/GermanYakimov/Text_tone_analyzer (date of access: 06.03.2020).