ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

УДК 004.08 DOI 10.25730/VSU.0536.19.015

Исследование методов анализа разнородных наборов данных

М. Л. Долженкова[1], Н. А. Долженкова[2], Г. А. Чистяков[3]

[1]кандидат технических наук, доцент кафедры электронных вычислительных машин, Вятский государственный университет. Россия, г. Киров. E-mail: maryid@mail.ru

[2]аспирант кафедры электронных вычислительных машин, Вятский государственный университет. Россия, г. Киров. E-mail: dolzhenkova.n.a@gmail.com

[3]кандидат технических наук, доцент кафедры электронных вычислительных машин, Вятский государственный университет. Россия, г. Киров. E-mail: gennadiychistyakov@gmail.com

Аннотация. В современной коммуникации большое место отводится общению в социальных интернет-сетях, которые позволяют им обмениваться информацией, самовыражаться, расширять круг интересов. Информация, находящаяся в сети, является полезной, но зачастую не может быть использована, так как является слабоструктурированной. В связи с этим актуальной является задача мониторинга, контроля и управления (администрирования) интернет-контентом, позволяющего своевременно анализировать большой объем разнородной информации (текстов, рекламы, изображений). В данной работе рассматриваются основные методы анализа текстовой и графической информации, предлагается комплексный подход к классификации оценки разнородных данных. В рамках проводимого исследования были выделены шкалы, по которым производится анализ тональности текстов. Практическое применение предлагаемого решения может быть найдено в области мониторинга социальных сетей для выявления девиантного поведения подростков.

Ключевые слова: обработка изображений, контент-анализ, тональность текста, девиантное поведение, мониторинг.

Введение. На данный момент 86% молодых людей в России в возрасте от 18 до 24 лет имеют аккаунт в социальной сети «В контакте» [1, с. 239], многие подростки проводят в интернете более четырех часов в день, а 8% учеников старших классов имеют признаки социальных девиаций, которые в том или ином виде могут быть обнаружены на их странице в социальной сети [4, с. 115].

Ввиду несовершенства существующих механизмов мониторинга специалисты, работающие в различных сферах, отмечают острую необходимость создания методов автоматизированного интернет-мониторинга.

Целью данной работы является анализ методов структурирования данных в существующих социальных интернет-сетях для выявления различных форм девиантного поведения, таких как суицидальные наклонности, употребление наркотических веществ, экстремистская деятельность и другие.

Материалы и методика исследований. Для обработки данных выбраны два наиболее подходящих способа обучения: продукционная экспертная система и нейронные сети.

Продукционная экспертная система – это инструментальный комплекс обработки данных, в которой экспертом формализуется большой объем информации (называемых базой знаний), в таком виде, что программный алгоритм может анализировать ее с помощью заданных экспертом правил.

Нейронные сети – математический аппарат, позволяющий на основе имеющихся данных получать прогнозы и классифицировать поступающую информацию. В случае разнородного набора данных имеет место следующая классификация:

1) текстовая информация;
2) графические изображения;
3) аудио- и видеоинформация.

Результаты исследований. Наиболее распространенным методом обработки текстовой информации является контент-анализ, предметом которого являются текстовые массивы данных. Метод ищет заданные экспертом маркеры, то есть конкретные слова, закономерности, темы; выявляет частоту их встречаемости в конкретном фрагменте и во всей выборке. Данный метод позволяет анализировать разноплановую информацию, имеющую диагностическое значение.

Для рассматриваемой предметной области такой информацией являются, прежде всего, текстовые сообщения на странице пользователя. Метод может помочь в выявлении конкретных слов, словосочетаний и хештегов, характерных для различных форм отклоняющегося поведения. Также стоит отметить, что при этом могут приниматься во внимание и эмотиконы – пиктограммы, отображающие эмоции.

Кроме того, важными характеристиками являются время и частота появления новых записей или «постов» на странице подростка, в том числе в динамике. Это позволяет косвенно отследить его режим сна и отдыха, потребность во внимании со стороны друзей в социальной сети. Так, например, количество друзей может указывать на широту круга общения, экстравертированность или интровертированность личности подростка.

К текстовой информации следует отнести названия аудио- и видеозаписей в соответствующих разделах аккаунта и на стене пользователя. В названиях определяющую роль играют не только слова-маркеры из текстовой базы, но и наиболее популярные в определенных субкультурах, тематических сообществах термины.

Важным этапом является выявление групп и публичных страниц, в которых состоит ребенок, так как именно группы отражают круг его интересов. Контент-анализ позволяет находить сообщества, посвященные различным девиантным темам, например, пропаганде употребления наркотиков или алкоголя. Еще одним методом анализа текстовой информации, ее позитивной или негативной окраски является оценка тональности текста [2, с. 8]. В рамках проводимого исследования были выделены шкалы, по которым производится анализ тональности.

1) Классификация по бинарной шкале. Использует два вида оценок: позитивная и негативная. Однако не всегда возможно однозначно определить, к какому классу необходимо отнести документ, так как он может содержать признаки обоих классов.

2) Классификация по многополосной шкале. Является усложнением предыдущего подхода – здесь градация полярностей включает в себя более двух вариантов. Например, часто по такой шкале просят оценить свое впечатление о фильмах, ресторанах, магазинах.

3) Системы шкалирования. Словам, связанным с полярными оценками, ставятся в соответствие числа по шкале от -10 до 10 (от самого отрицательного к самому положительному). Далее текст исследуется при помощи методов обработки естественного языка, а затем выделенные из текста объекты анализируются с целью их понимания.

4) Субъективность/объективность. Задача – установить является текст субъективным или объективным, что сложнее, чем классификация полярности, так как субъективность может зависеть от контекста, а объективный текст может содержать субъективные мнения (например, статья, где цитируется чье-то мнение).

Помимо выбора системы шкал был произведен анализ подходов обработки текста в целом. Было выделено три группы подходов:

a) основанные на словарях;

b) основанные на заданных правилах;

c) основанные на методах машинного обучения.

В первом случае используются специализированные тональные словари, представляющие из себя список терминов и соответствующих им значений эмоциональной окраски. В процессе анализа каждому слову присваивают соответствующее значение тональности (при наличии такого слова в словаре), а затем вычисляют общую тональность текста (вычисление можно выполнять разными способами, например, как среднее арифметическое всех значений). При обработке графической информации более оптимально применять метод многоуровневого анализа.

На первом уровне выполняется поиск текста на изображении, который затем обрабатывается по правилам текстовых маркеров. На втором уровне отслеживаются изображения определенной цветовой гаммы, например, содержащие в основном темные или тусклые цвета. На практике на основе наличия большого количества изображений депрессивной гаммы можно выявить подавленное настроение пользователя. Третий уровень – это анализ примитивов.

Работа на каждом из уровней основывается на двух последовательных шагах – классификация изображения и его оценка в контексте группы.

Среди известных подходов к построению классификации выделяют следующие:

1) подход Фу (Fu) [3, с. 7];
2) подход Пала (Pal) [3, с. 10];
3) подход Скарбека и Кошана (Skarbek и Koschan) [7, p. 23];
4) подход Лючиса и Митра (Lucchese и Mitra) [6, p. 114].

Каждый из рассмотренных подходов имеет ряд существенных недостатков, что не позволяет применить его в чистом виде. В связи с этим предлагается комплексный подход, состоящий из следующих шагов.

1. Определение свойств, на основе которых выполняется сегментация (разрывность или сходство низкоуровневых признаков).
2. Выбор стратегии обработки изображения (последовательная или параллельная).
3. Определение типа изображения (цветное или полутоновое), к которому применяются алгоритмы сегментации.

Наиболее часто в задачах распознавания и идентификации изображений используются классические нейросетевые архитектуры (многослойный персептрон, сети с радиально-базисной функцией и др. [5, p. 42]), но применение классических нейросетевых архитектур в задачах распознавания имеет ряд недостатков:

a) для изображений большой размерности существенно возрастает количество нейронов сети;

b) большое количество параметров требует большего объема обучающей выборки, увеличивает время и вычислительную сложность процесса обучения;

c) для повышения эффективности работы системы желательно применять несколько нейронных сетей (обученных с различными начальными значениями синаптических коэффициентов и при разном порядке предъявления образов), что увеличивает вычислительную сложность и время решения задачи;

d) отсутствует инвариантность к изменениям масштаба изображения, ракурсов съемки камеры и других геометрических искажений входного сигнала.

Для решения данной проблемы в настоящий момент сверточные нейронные сети, обеспечивающие частичную устойчивость к изменениям масштаба, смещениям, поворотам, смене ракурса и прочим искажениям. Для повышения устойчивости часто используется масштабирование.

Выводы. В результате проведенных исследований была выявлена необходимость разработки новых методов выявления в слабоструктурированных данных устойчивых закономерностей, связанных в том числе с поведением детей и подростков в социальных сетях. Предлагаемый подход является комплексным и сопоставимым по эффективности с ручным мониторингом социальных сетей профильным специалистом.

Проведена теоретическая научно-исследовательская работа для ряда семантических тезаурусов. Для проведения практической части работы заключены партнерские соглашения с тремя школами города Кирова.

Список литературы

1. *Гудакова Л. В., Чернышев Е. А., Шереметова А. И.* Влияние деструктивного интернет-контента на формирование девиантного поведения у подростков // Образование и наука в современных реалиях : сборник материалов V Международной научно-практической конференции / редкол.: О. Н. Широков [и др.]. 2018. С. 239–242.

2. *Котельников Е. В.* Метод анализа тональности текстов TextJSM // Научно-техническая информация. Сер. 2. 2018. № 2. С. 8–20.

3. *Садыхов Р. Х., Ваткин М. Е.* Модифицированный алгоритм обучения РБФ-сети для распознавания рукописных символов // Идентификация образов. 2001. Т. 1. № 3. С. 7–16.

4. *Щетинина Е. В.* Работа киберлаборатории как фактор профилактики экстремистских и террористических проявлений в сети Интернет // Вестник Южно-Уральского государственного университета. Право. 2018. Т. 18. № 1. С. 115–119.

5. *Feraud R., Bernier O., Viallet J., Collobert M.* A fast and accurate face detector based on neural networks // Trans. Pattern Anal. Machine Intelligence. 2002. V. 3. № 23. P. 42–53.

6. *Lin S., Kung S., Lin L.* Face recognition detection by probabilistic decision-based neural network // Trans. Neural Networks. 1997. V. 8. № 1. P. 114–132.

7. *Rowley H. A., Baluja S., Kanade T.* Neural network-based face detection // Pattern Anal. Mach. Intell. 2000. V. 5. P. 23–38.

Research of methods of analysis of heterogeneous data sets

M. L. Dolzhenkova[1], N. A. Dolzhenkova[2], G. A. Chistyakov[3]

[1]PhD of technical sciences, associate professor of the Department of electronic computing machines, Vyatka State University. Russia, Kirov. E-mail: maryid@mail.ru

[2]post-graduate student of the Department of electronic computing machines, Vyatka State University. Russia, Kirov. E-mail: dolzhenkova.n.a@gmail.com

[3]PhD of technical sciences, associate professor of the Department of electronic computing machines, Vyatka State University. Russia, Kirov. E-mail: gennadiychistyakov@gmail.com

Abstract. In modern communication, a great place is given to communication in social Internet networks, which allow them to exchange information, express themselves, expand the range of interests. The information in the network is useful, but often can not be used because it is poorly structured. In this regard, the task of monitoring, control and management (administration) of Internet content, allowing timely analysis of a large amount of heterogeneous information (texts, advertising, images) is relevant. This paper discusses the main methods of analysis of text and graphic information, offers an integrated approach to the classification of heterogeneous data estimates. Within the framework of the study, the scales on which the tonality of texts is analyzed were identified. Practical application of the proposed solution can be found in the field of monitoring of social networks to identify deviant behavior of adolescents.

Keywords: image processing, content analysis, text tonality, deviant behavior, monitoring.

References

1. *Gudakova L. V., Chernyshev E. A., Sheremetova A. I. Vliyanie destruktivnogo internet-kontenta na formirovanie deviantnogo povedeniya u podrostkov* [Influence of destructive Internet content on the formation of deviant behavior of adolescents] // *Obrazovanie i nauka v sovremennyh realiyah: sbornik materialov V Mezhdunarodnoj nauchno-prakticheskoj konferencii* – Education and science in modern realities: proceedings of the V International scientific-practical conference / ed. board: O. N. Shirokov [et al]. 2018. Pp. 239–242.

2. *Kotel'nikov E. V. Metod analiza tonal'nosti tekstov TextJSM* [Method of text tonality analysis TextJSM] // *Nauchno-tekhnicheskaya informaciya* – Scientific and technical information. Ser. 2. 2018. No. 2. Pp. 8–20.

3. *Sadyhov R. H., Vatkin M. E. Modificirovannyj algoritm obucheniya RBF-seti dlya raspoznavaniya rukopisnyh simvolov* [Modified learning algorithm of RBF network for recognition of handwritten symbols] // *Identifikaciya obrazov* – Identification of the images. 2001. Vol. 1. No. 3. Pp. 7–16.

4. *Shhetinina E. V. Rabota kiberlaboratorii kak faktor profilaktiki ekstremistskih i terroristicheskih proyavlenij v seti Internet* [Work of the cyber laboratory as a factor of prevention of extremist and terrorist manifestations on the Internet] // *Vestnik Yuzhno-Ural'skogo gosudarstvennogo universiteta. Pravo* – Herald of the South Ural State University. Law. 2018. Vol. 18. No. 1. Pp. 115–119.

5. *Feraud R., Bernier O., Viallet J., Collobert M.* A fast and accurate face detector based on neural networks // Trans. Pattern Anal. Machine Intelligence. 2002. V. 3. № 23. Pp. 42–53.

6. Lin S., Kung S., Lin L. Face recognition detection by probabilistic decision-based neural network // Trans. Neural Networks. 1997. V. 8. № 1. Pp. 114–132.

7. *Rowley H. A., Baluja S., Kanade T.* Neural network-based face detection // Pattern Anal. Mach. Intell. 2000. V. 5. Pp. 23–38.