УДК 004.81/.032.26

МОДЕЛЬ ВЛИЯНИЯ ТЕКСТОВЫХ КОММЕНТАРИЕВ НА ПСИХОЭМОЦИОНАЛЬНОЕ СОСТОЯНИЕ АУДИТОРИИ ИНТЕРНЕТ-СМИ

Монахов Ю.М., Маков Е.О.

ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича
и Николая Григорьевича Столетовых», Владимир, e-mail: unklefck@gmail.com

В данной статье рассматриваются методы автоматического определения эмоциональной составляющей текста или его тональности. Изложен алгоритм работы системы для комментариев материалов интернет-СМИ на русском языке, в основе которого лежит анализ психоэмоциональных показателей и машинная синестезия. В работе впервые предложен метод определения тональности, основанный на выделении психоэмоциональной структуры текста комментария. В основе данного метода лежат самоорганизующиеся карты Кохонена, которые обеспечивают выявление скрытых и кластеризацию известных признаков объекта методом неконтролируемого обучения. Для преобразования многомерного пространства входов нейронов сети используется метод сжатия без потери информативности – неотрицательная матричная факторизация (NMF). Конечным этапом преобразований является получение изображения с целью выявления тональности исходного текста, для классификации используется метод опорных векторов с нелинейным ядром.

Ключевые слова: Интернет-СМИ, информационно-психологическое воздействие, эмоций, тональность, метод определения тональности, психоэмоциональный дифференциал, машинная синестезия

A MODEL OF INFLUENCE OF TEXT COMMENTS ON THE PSYCHO-EMOTIONAL STATE OF ONLINE MEDIA AUDIENCE

Monakhov Yu.M., Makov E.O.

Vladimir State University named after Alexander and Nickolay Stoletovs, Vladimir;
e-mail: unklefck@gmail.com

This article concerns the methods of automatically determining the emotional component in the text and describes the algorithm of the system for text comments of online media in the Russian language, which is based on emotional differentials and engine synesthesia. This work first proposes the method of determining the tone, based on allocation of psycho-emotional structure of text comment. The basis of this method is self-organizing Kohonen's map, which provides identification of concealed features and known features clustering of the object by the method of unsupervised learning. To convert the multidimensional space of inputs of processing elements is used the method of compression without information loss – non-negative matrix factorization (NMF). The final stage of the transformation is getting an image to determine the tone of the source text, for classification support vectors machine with nonlinear kernel is used.

Keywords: online media, information and psychological effects, emotions, the tone, the method of determining the tone, the emotional differential, engine synesthesia

В настоящее время растет количество ресурсов, в которых пользователи интернета могут получать новости и свободно высказывать свое мнение о них. В связи с этим существует и продолжает расширяться спектр атак на информационно-психологическую безопасность личности человека. Защита человека от психологического воздействия является актуальной проблемой, так как методы, используемые для изменения состояния информационной среды и условий жизнедеятельности, способствуют нарушению ценностей, адаптивности функционирования и развития. Через комментарии к статьям СМИ можно воздействовать на объект посредством написания негативных сообщений, которые отрицательно влияют на психологическое и эмоциональное состояние человека [3].

Исходя из вышесказанного, цель работы определена как разработка модели

влияния текстовых комментариев на психоэмоциональное состояние аудитории интернет-СМИ.

В соответствии с целью мы определили объект и предмет исследования. Объектом данного исследования являются интернет-СМИ, а предметом исследования – методы автоматизированной оценки психоэмоциональной тональности и классификация комментариев источника аудитории СМИ.

В работе под определением Интернет-СМИ мы будем рассматривать множество сайтов в информационно-телекоммуникационной сети Интернет, деятельность которых связана с периодическим распространением массовой информации под постоянным наименованием (названием).

Исходя из определения, интернет-ресурсы могут быть причислены к профессиональной журналистике, если целевые функции отвечают требованиям, предъявля-

емым выборными демократиями к прессе, которая должна информировать граждан, а также способствовать их свободам и самоуправлению. Интернет-СМИ описываются с помощью формулы Лассуэлла по пяти главным классификационным признакам: канал, аудитория, коммуникатор, контент, эффекты [7, 8].

Для дальнейшей классификации медийного поля интернета были взяты следующие типы ресурсов:

- индексирующие и классифицирующие информацию по категориям (index and category sites) – Yahoo!;
- информационные агентства (mainstream news sites) – CNN, BBC, MSNBC, Интерфакс, ИТАР-ТАСС, РИА «Новости» и др.;
- метамедийные сайты (meta and comment sites) – это ресурсы экспертного характера, посвященные исключительно СМИ и журналистике;
- сайты, предназначенные для обмена информацией и для дискуссий (share and discussion sites) [2].

Существующие методы

В последние годы ведется много исследований в области определения тональности текстов. Данные разбиваются на две категории (положительные и отрицательные), либо на три (положительные, отрицательные и нейтральные). Много работ посвящено тоновой классификации комментариев на продукты и фильм, анализу блогов и новостей. Были исследованы длинные тексты, принадлежащие некоторой заранее определенной предметной области. Для исследований разрабатывались тренировочные корпуса текстов, обладающие следующими параметрами:

- корпуса отзывов с вручную предоставленными потребителями оценками;
- узкотематические корпуса отзывов;
- корпуса общезначимых новостей.

В статье [6] авторы разработали систему определения тональности с помощью глубокого сантмент-анализа, на основе лингвистического анализа текста на естественном языке. Как показывают результаты, с помощью этого метода можно достичь достаточно высокой точности на текстах новости СМИ. Но одной из причин ошибок является ограниченность используемого эмотивного пространства: часть лексики не попадает в пространство хорошо-плохо. Разграничение текста только на негативный и позитивный при разметке словаря сужает способы обработки данных и не решает вопрос, который лежит в области понимания и восприятия информации мозгом человека.

В работах [4, 5] авторы получили результаты для ДСМ-метода решения задачи определения тональности и сравнили его с результатами частотного классификатора и с методом опорных векторов. Результаты, полученные на тестовой коллекции с использованием ДСМ-метода, лучше результатов базового классификатора, на 4% лучше результатов SVM-классификатора без подбора оптимального числа характеристик.

Промышленные системы автоматической обработки текста используют следующую схему обработки текстов, состоящую из трех этапов:

- морфологический анализ (определение морфологических характеристик каждого слова – часть речи, падеж, склонение, спряжение и т.д.) и морфемный анализ (приставка, корень, суффикс и окончание);
- синтаксический анализ;
- задачи семантического анализа (поиск фрагментов, формализация, реферирование и т.д.).

Разработка модели

В настоящее время разработано немало количество различных методов, используемых для решения сходных задач. Ниже мы рассмотрим некоторые методы и их адаптацию к данной работе.

Поиск по неоднородным данным со сложной признаковой структурой может проводиться различными методами. Наиболее часто используемым является мера информативности термина. TF-IDF – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален количеству употребления этого слова в документе и обратно пропорционален частоте употребления слова в других документах коллекции.

Мера TF-IDF часто используется в задачах анализа текстов и информационного поиска, например, как один из критериев релевантности документа поисковому запросу, при расчёте меры близости документов при кластеризации [12, 15].

TF – частота слова. Оценивается важность слова в пределах отдельного документа.

$$tf(t, d) = n_t / \sum_k n_k, \quad (1)$$

где n_t есть число вхождений слова в документ, а в знаменателе – общее число слов в данном документе.

IDF – обратная частота документа. Обратное значение частоты, с которой слово встречается в корпусе документов. Учёт IDF понижает вес часто употребляемых слов.

$$idf(t, D) = \log \left(\frac{|D|}{|\{d_i \in D | t \in d_i\}|} \right), \quad (2)$$

где $|D|$ – количество документов в корпусе, $|\{d_i \in D | t \in d_i\}|$ – количество документов, в которых встречается термин t [13, 16].

Методы, основанные на модели векторного пространства, используют формируемый заранее вектор признаков документа, включающий в себя список наиболее значимых терминов. Дальнейшее сравнение документов проводится путем оценки сходства между полученными векторами с использованием меры косинусного сходства. Косинусное сходство – это метод, основанный на модели векторного пространства, который использует формируемый заранее вектор признаков документа, включающий в себя список наиболее значимых терм, где сравнение документов проводится путем оценки сходства между полученными векторами с использованием измерения косинуса угла между ними [11].

Если даны два вектора признаков, A и B , то косинусное сходство может быть представлено, используя скалярное произведение и норму.

$$\cos(x, y) = \frac{|xy|}{\sqrt{|x| * |y|}}. \quad (3)$$

Для задачи кластеризации данных и выделения новых признаков объекта были взяты карты Кохонена. Самоорганизующиеся карты – это тип нейронных сетей, которые используют неконтролируемое обучение. Обучающее множество таких моделей состоит из значений входных переменных, в процессе обучения нет сравнивания выходов нейронов с эталонными значениями. Такая сеть способна понимать структуру данных [10, 14].

Приложения, основанные на снижении размерности данных, используют метод факторизации, который позволяет найти компактный способ описания данных, со-

храняющий большую часть содержащейся в них информации. Использование низкоразмерных представлений позволяет экономить вычислительные ресурсы и уменьшить число операций в случаях, когда оно зависит от размерности данных. Одним из таких методов факторизации матриц являются неотрицательные матричные разложения.

Задача неотрицательного матричного разложения (NMF) ставится как оптимизационная: необходимо найти неотрицательные факторы, доставляющие минимум некоторому функционалу потерь. Выбор данной функции оказывает большое влияние на получаемое решение. В разных прикладных областях для построения неотрицательного матричного разложения используются разные функции потерь: в тематическом моделировании используется дивергенция Кульбака-Лейблера, в биологических приложениях – норма Фробениуса, в анализе аудиозаписей – дивергенция Итакуры – Сайто, в некоторых задачах машинного зрения – метрика EMD [9].

Подход к решению

Вместо традиционных методов определения тональности текста, таких как классификация векторов семантических дифференциалов или ДСМ-метод, в нашем подходе используется многоэтапное преобразование данных, основанное на психоэмоциональных показателях, с их дальнейшей классификацией.

Мы выделили четыре стадии преобразования данных (рисунок):

- выделение слов из комментария и их лемматизация;
- вычисление для каждого уникального слова вектора эмоциональных показателей tf-idf;
- инициализация самоорганизующейся карты кохонена и ее обучение на полученных из предыдущего этапа векторов;
- снижение размерности весовых показателей синапсов самоорганизующейся карты неотрицательным матричным разложением до двухмерного представления.

Схема преобразования данных

Начальным этапом преобразования данных является выделение слов из текста, где каждое слово будет преобразовано в нормальную форму.

В данной работе мы учитываем эмоциональную окраску каждого слова с целью определения психоэмоциональной значимости относительно окраски корпуса текстов. В результате была произведена корректировка формулы вычисления TF-ID. Каждому слову в словаре будет назначаться 11-мерный вектор, где каждый его элемент, характеризует одну из базовых эмоций: Удивление, Печаль, Гнев, Отвращение, Презрение, Горе-страдание, Стыд, Интерес-волнение, Вина, Смущение, Радость (классификация эмоций по К. Изарду [1]).

После дополнений формулы, она приобрела следующий вид:

$$tf - idf = tf * idf, \quad (4)$$

где множители принимают следующий вид:

$$tf(t, e, d) = \frac{k * t_e}{10 * n}, \quad (5)$$

где k – количество раз использования термина в тексте d , t_e – коэффициент e -ой эмоции термина t , n – количество слов в документе.

$$idf(t, e, d, D) = \log \frac{10 * |D|}{\sum_{t \in d \mid d \in D} d_e}, \quad (6)$$

где t – термин, e – номер эмоции, d – текст, D – корпус исследуемых текстов, $\sum_{t \in d \mid d \in D} d_e$ – сумма e -ых эмоциональных коэффициентов всех текстов, которые содержат термин t .

Механизм машинной синестезии изображения с целью выявления скрытых эмоциональных характеристик слов предлагаем основывать на картах Кохонена. В начале инициализируем нейронную сеть с весовыми коэффициентами равными обратным значениям коэффициентов сети, обученной на всем эталонном корпусе слов. Обратное значение вычисляется как $1/w_{ij}$, где w_{ij} – вес i -го нейрона j -го синапса. Далее на вход сети поочередно поступают психоэмоциональные вектора слов, содержащихся в тексте. Результатом данного этапа является кластеризация данных в многомерном пространстве по показателю признаков.

Далее, полученное многомерное пространство будет сжато NMF методом. Оптимальность функции потерь в конкретной прикладной задаче зависит от структуры шума, содержащегося в данных. В данном исследовании была использована норма Фробениуса, для преобразования дан-

ных синаптических весов нейронной сети в двухмерное изображение.

Итогом работы алгоритма является преобразование изображения в n -мерный вектор, где $n = h * d$, h – высота изображения, d – ширина изображения, который подается на вход классификатору. Классификатор основан на методе опорных векторов, ядро представлено сигмоидальной функцией. Классификатор обучается на размеченной эталонной выборке текстов, которые были получены путем социально-психологического опроса. Тексты разделены на три класса: негативные, нейтральные и позитивные.

Выводы

Практическая значимость исследования заключается прежде всего в создании методики психоэмоциональной оценки текстовых сообщений, автоматизация которой в дальнейшем позволит включить её в качестве модуля в состав автоматизированной системы фильтрации пропагандистских, агитационных, рекламных и прочих материалов, вызывающих негативное психоэмоциональное воздействие на личность человека. Это позволит сократить время обработки процедуры анализа, сократить количество экспертов, работающих на данном этапе процедуры, а также значительно повысить объёмы анализируемой информации. И таким образом, общий процент фильтрации сообщений, не удовлетворяющих требованиям пользователя, будет повышаться.

Список литературы

1. Изард К. Эмоции человека. – М.: Директ-Медиа, 2008. – 954 с.
2. Интернет СМИ: Теория и практика: Учеб. пособие для студентов вузов / И73 Под ред. М.М. Лукиной. – М.: Аспект Пресс, 2010. – 348 с.
3. Кабаченко Т.С. Методы психологического воздействия. Учебное пособие. – М.: Пед. общ. России, 2000. – 544 с.
4. Клековкина М.В., Котельников Е.В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // Труды XIV Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – 2012. – С. 118–123.
5. Котельников Е.В. Опыт применения ДСМ-метода для определения тональности текста // Тринадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2012 (16–20 октября 2012 г., г. Белгород, Россия): Труды конференции. Т. 1. – Белгород: Изд-во БГТУ, 2012. – 377 с.
6. Пазельская А.Г., Соловьев А.Н. Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). – М.: Изд-во РГТУ. – 2011. – № 10. – С. 17.

7. Подставко Е.Н. Классификации сетевых изданий Рунета // Научные ведомости БелГУ. Серия: Гуманитарные науки. – 2011. – № 6 (101). – С. 264–269.
8. Сюндюков Н.К. Интернет-СМИ и особенности их функционирования // Управленческое консультирование. – 2014. – № 12 (72). – С. 180–191.
9. Рябенко Е.А. Мультипликативный метод неотрицательного матричного разложения с АБ-дивергенцией и его сходимость // Машинное обучение и анализ данных. – 2014. – Т. 1, № 7. – С. 800–816.
10. Хайкин С. Нейронные сети: полный курс, 2-е издание. – М.: Издательский дом Вильямс, 2008 – С. 1104.
11. Чубукова И.А. Data Mining. – 2008. – 326 с.
12. Chum O. et al. Near Duplicate Image Detection: min-Hash and tf-idf Weighting // BMVC. – 2008. – Т. 810. – P. 812–815.
13. Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. – Carnegie-mellon univ pittsburgh pa dept of computer science, 1996. – № CMU-CS-96-118.
14. Kohonen T. Self-Organizing Maps (Third Extended Edition). – New York, 2001. – 501 p.
15. Ramos J. et al. Using tf-idf to determine word relevance in document queries // Proceedings of the first instructional conference on machine learning. – 2003. – Т. 242. – P. 133–142.
16. Zhang W., Yoshida T., Tang X. A comparative study of TF* IDF, LSI and multi-words for text classification // Expert Systems with Applications. – 2011. – Т. 38, № 3. – P. 2758–2765.