НОВЫЕ МЕДИА

Бодрунова Светлана Сергеевна, доктор политических наук, профессор кафедры менеджмента массовых коммуникаций, Институт «Высшая школа журналистики и массовых коммуникаций», Санкт-Петербургский государственный университет, Санкт-Петербург, Россия; e-mail: s.bodrunova@spbu.ru

КРОСС-КУЛЬТУРНЫЙ ТОНАЛЬНЫЙ АНАЛИЗ ПОЛЬЗОВАТЕЛЬСКИХ ТЕКСТОВ В ТВИТТЕРЕ[1]

В статье представлен обзор современных исследований в области тонального анализа естественного языка (sentiment analysis, сентимент-анализ) в сети Интернет, в том числе на платформе микроблогов Twitter (Твиттер). Описаны направления современных исследований тональности твитов, методологические, технические и иные проблемы при анализе наборов данных, ограничения существующих подходов. Выделены группы современных исследований на основании нескольких критериев: наличие машинного обучения, единица анализа, объект исследования. Показано, что наиболее крупной задачей современного тонального анализа текстов остается создание многоязычных инструментов, которые позволяли бы сравнительный анализ тональности высказываний и целых дискуссий в межстрановой и кросскультурной перспективе. Оценены существующие инструменты кросскультурного сентимент-анализа. Предложена концепция для разработки оригинального инструмента для анализа тональности высказываний в Твиттере на нескольких языках.

Ключевые слова: *тональный анализ, сентимент, Твиттер, вычислительные методы в социальных науках, кросс-культурный тональный анализ.*

Svetlana S. Bodrunova, Doct.Sci. in Political Science, Professor at Department of Mass Media Management, Institute 'School of Journalism and Mass Communications', St.Petersburg State University, St.Petersburg, Russia; e-mail: s.bodrunova@spbu.ru

CROSS-CULTURAL SENTIMENT ANALYSIS OF USER TEXTS ON TWITTER

The paper reviews today's most successful approaches to sentiment analysis of massive datasets of user-generated texts, including those from Twitter. We define today's most developed areas of sentiment studies and their limitations, as well as methodological, technological, and other challenges that sentiment analysis faces in its variations across the world. We also group the existing research into clusters based on several criteria, including presence/absence of machine learning, unit of analysis, and object of study. We show that the creation of cross-cultural

multilingual sentiment analysis and tools for it is a major task that today's sentiment studies face; such tools would allow detecting sentiment across a range of languages and cultures. We assess the existing tools for multi-lingual sentiment analysis and suggest a conceptual framework for future studies of sentiment in different language domains of Twitter.

Key words: *sentiment analysis, sentiment, Twitter, computational social science, cross-cultural sentiment analysis.*

DOI: 10.30547/vestnik.journ.6.2018.191212

Введение: тональный анализ естественного языка и текстов в сети Интернет

Современные пользовательские дискуссии в сети Интернет содержат важную социальную информацию, которую можно выявлять и изучать междисциплинарными методами. К такой информации могут относиться социальные и политические аттитюды, культурные коды, паттерны межличностного общения и распространения знаний, информация о групповых интересах и конфликтных настроениях. Картирование онлайн-дискуссий и изучение их контента имеет прогностический и антиконфликтный потенциал.

Для адекватного картирования сетевой дискуссии необходимо выявить взаимосвязь между тремя компонентами: 1) структурой дискуссии (характером связей между ее узлами, характеристиками ведущих пользов ателей, наличием внутренних разломов и кластеров); 2) тематикой дискуссии и ее отдельных фрагментов; 3) тональностью дискуссии относительно обсуждаемого объекта (дискриминируемой социальной группы, политического актора, коммерческого продукта и т. п.). Но сетевые дискуссии остаются малоизученными с точки зрения связи структуры дискуссии с характером обсуждения, в т. ч. с эмоциональной составляющей пользовательской коммуникации.

Помочь в изучении эмоций пользователей способен тональный анализ (*sentiment analysis*, сентимент-анализ) – область исследований на базе машинных методов обработки естественного языка (*natural language processing*), посвященная выявлению тональности высказывания (эмоционального отношения/оценки, чаще всего позитивной/негативной) – либо по отношению к обсуждаемому в тексте объекту, либо в тексте в целом.

Также для обозначения данной области исследований применяются термины *анализ субъективности* (*subjectivity analysis*) и *глубин-*

ный анализ мнений (*opinion mining*; Dave et al., 2003). Большинство исследователей рассматривают сентимент-анализ как метод классификации текстов/документов. Следует различать классификацию по тональности (*document sentiment classification*) – и классификацию по объективности/субъективности (*subjectivity classification*) и только затем по тональности (только для субъективных документов).

Opinion mining – комплексная методология, включающая три этапа: 1) определение релевантных корпусов текстов; 2) выявление среди найденного корпуса текстов, содержащих мнение (*opinionated content*); 3) определение тональности высказывания авторов текстов и кластеризация документов согласно выявленной тональности. Именно такой подход следует рассматривать в качестве базового для работы с короткими пользовательскими текстами (Liu, 2010).

Первая волна взрывного интереса к тональному анализу и анализу субъективных мнений в пользовательских текстах относится к началу 2000-х гг. (Pang, Lee, 2008). С тех пор сложилось поле исследований тональности англоязычных текстов разного происхождения – настолько разветвленное и разнообразное, что позволяет даже говорить о его расплывчатости (Kan, 2012). Развитию конкурентных методов тонального анализа способствовали международные воркшопы. Так, команда под руководством Накова и Розенталя в рамках воркшопов *SensEval* и *SemEval* предложила для развития методов тонального анализа пять больших заданий (*tasks*), в которых разные группы ученых со всего мира применяли различные методики сентимент-анализа для решения одной и той же задачи (Nakov et al., 2016; Rosentahl, Farra, Nakov, 2017). Проводились подобные воркшопы и в России.

Современные подходы и методы в области тонального анализа

Основой тонального анализа является идея об автоматической классификации текстовых документов по *тону высказывания* (сентименту), то есть разделение документов по авторской эмоции – на группы условно негативных, условно позитивных и условно нейтральных текстов с помощью автоматизированных методов. Основанием для разделения текстов является лексический уровень (наборы слов с присвоенными негативными или позитивными значениями), но многие методы учитывают грамматический строй речи, связи между лексическими единицами, а также иронию и многозначность.

Кластеризация текстов по тону может применяться в т. ч. для наборов традиционных медиатекстов и текстов из социальных медиа. Тональный анализ позволяет решать массу прикладных задач текстового анализа, например выделять радикально эмоциональные тексты, потенциально содержащие призывы к насилию и розни, и описывать настроения журналистов и пользователей по поводу событий, персон, конфликтов. Сегодня именно социальные медиа становятся основным объектом тонального анализа в *computational social science*, т. е. социальных науках на основе вычислительных методов.

Методы тонального анализа можно разбивать на подгруппы по разным признакам.

1) По наличию машинного обучения:

- методы, основанные на знании (*knowledge-based approaches*), то есть на изучении конкретного корпуса текстов и лексики в нем;
- методы, основанные на правилах и шаблонах (Schuller, Knaup 2011; Клековкина, Котельников, 2012), закрепленных в языке;
- методы машинного обучения без учителя (*non-supervised machine learning approaches*);
- методы машинного обучения с частичным/полным присутствием учителя (*semi-supervised/supervised machine learning approaches*) (Shalunts, Backfield, 2016).

Как правило, это разделение описывается как разделение между методами на основе лексиконов и методами машинного обучения (Pang et al., 2002; Goncalves et al., 2013). Методы на основе лексиконов, как указано выше, делятся на словарные методы и методы на основе конкретных корпусов текстов (статистические и семантические) (Serrano-Guerrero et al., 2015). Размеченные вручную и/или с помощью словарей лексиконы/тезаурусы содержат информацию как о словах и фразах, так и о типах и силе связей между ними.

Методы машинного обучения – без учителя и с учителем – предполагают обучение машины распознавать эмоции пользователей на основе размеченной выборки текстов, закодированной вручную, с помощью кодировщиков (так называемой эталонной коллекции). Алгоритм обучения и классификации текстов принято называть классификатором. Разработаны различные типы классификаторов, которые делятся на:

- классификаторы на основе древа решений (*decision tree classifiers*) (Serrano-Guerrero et al., 2015);

- линейные классификаторы: на основе опорных векторов (*support vector machines, SVM*) (Barbosa, Feng, 2010; Pak, Paroubek, 2010; Kontopoulos et al., 2013), на основе нейронных сетей;

- классификаторы на основе правил и шаблонов;

- вероятностные классификаторы: наивный Байесов классификатор, Байесовы сети, классификатор на основе метода максимальной энтропии.

Многие исследования строятся на сочетании лексикон-ориентированных методов и методов машинного обучения. Улучшение *SVM*-классификаторов составляет отдельную научную проблему (Sharma, Dey, 2013). Считается, что для малоструктурированных текстов с высоким уровнем шума – таких, какие и бывают в соцсетях, – пригодны модели на основе *мешка слов* (*bag-of-words approach*; Yessenov, Misailovic, 2009), то есть методы вероятностной оценки со-встречаемости слов. Иногда для векторных классификаторов и классификаторов на основе *мешка слов* предлагается учитывать веса слов в текстах, в т. ч. их частоту (Tellez et al., 2017).

2) По избранной единице анализа выделяют подходы, имеющие единицей классификации или весь документ, или фрагменты текста – синтагмы и отдельные термины, где под термином понимается слово или словосочетание, свойственное тексту (Kouloumpis et al., 2011). Тональный анализ может проводиться на уровне леммы, синсета (комплекта связанных между собой семантическими связями эмоционально окрашенных лексем), синтагмы/предложения (*sentence*), всего документа (Liu, 2012; Cruz et al., 2014).

Отдельное место в обсуждении методологии уделяется способам репрезентации и понимания термина (в указанном выше значении). Одним из таких способов является выделение в текстах так называемых *n*-грамм – формальных фрагментов в несколько лексических единиц, где ключевое слово окружено лексемами с обеих сторон (вида $1\pm n$). Так можно оценить, в каком тональном *окружении* находится ключевое слово. Чаще всего для анализа используются униграммы и биграммы, то есть фрагменты самой малой длины – в три и пять слов. Еще одним важным дополнением к знанию о словарном составе исследуемого текста является автоматизированная расстановка (*labeling*) частей речи.

3) По объекту анализа (Loukashevitch et al., 2015; Pontiki et al., 2016):

- общий сентимент-анализ (для определения тональности всего текста);

- подходы, ориентированные на объект (*object-related/entity-related approaches*), т. е. направленные на выявление отношения к некоему объекту (политику, компании, событию и т. д.). Может использоваться для мониторинга репутации, отслеживания настроения избирателей и т. д.;

- подходы, ориентированные на аспект/характеристику такого объекта (*aspect-related/feature-based approaches*). Технически мало чем отличаются от объект-ориентированного сентимент-анализа, но ориентированы на идентификацию отношения пользователей к отдельным характеристикам объекта. Например, аспект-ориентированный тональный анализ может применяться к оценке мнений комментаторов в рекомендательных сервисах, где для объекта (например, кинотеатра) оцениваются стоимость билетов, транспортная доступность, удобство кресел, звук, экраны и т. д.

Иные методы включают разные способы улучшения опознаваемости эмоциональных высказываний путем комбинации тонального анализа с другими методами анализа текста. Это в том числе:

- кросс-доменное улучшение качества тонального анализа (Ponomareva, Thelwall, 2012), то есть попытки переноса методов, разработанных для какой-то одной тематики (скажем, рецензий на книги) в другие тематические области. Это, скорее, не метод, а подход, включающий целый комплекс методик, в т. ч. некоторые из перечисленных ниже;

- графовый тональный анализ (*graph-based approach to sentiment analysis*) (Усталов, 2012; Ponomareva, Thelwall, 2012). Это комплекс методик, основанный на вероятностных алгоритмах оценки совстречаемости слов с уже известным (например, словарным) сентиментом и слов, сентимент которых неизвестен и выявляется путем многократного перебора вероятностей их нахождения рядом с размеченными словами. При этом каждому слову с неизвестным сентиментом в каждой итерации присваивается значение, равное сумме окружающих слов – как размеченных, так и не размеченных; с каждой итерацией значение для каждого слова уточняется и объем знания о сентименте в корпусе текстов растет. Алгоритм останавливается, когда достигает так называемой точки конвергенции, когда значения для всех слов определены с приемлемой (установленной заранее) точностью;

- семантический тональный анализ (Saif et al., 2012), суть которого в добавлении семантической категории к объекту или классу объектов (например, продукт *Apple* – к *iPhone*). Эмоции пользова-

телей оцениваются не только по отношению к объекту, но и по отношению к семантической категории;

- сочетание тонального анализа с методами вероятностной кластеризации текстов, например тематическим моделированием (Li et al., 2010; Ohmura et al., 2014), и оценка сентимента по отдельным кластерам, например отдельным темам внутри набора данных;

- применение методов дискурс-анализа к разметке текстов для тональной оценки (Heerschop et al., 2011);

- тональный анализ с помощью нейронных сетей (*neural-networks-based sentiment analysis)* (Dos Santos, Gatti, 2014). Этот метод привлекателен тем, что его результаты не всегда соответствуют ожиданиям исследователей и позволяют высококачественно выполнить задачу, но результаты применения нейронных сетей не всегда поддаются интерпретации с алгоритмической и каузальной точек зрения.

Оценка качества анализа. Традиционно оценку качества тонального анализа проводят двумя способами. Во-первых, качество оценивают путем сравнения результатов автоматизированного анализа с данными эталонной коллекции. Еще один путь – ручное кодирование эталонной коллекции, разработка лексикона на ее основе, его применение для оценки тональности текстов и затем повторная оценка кодировщиками точности автоматического присвоения тональности.

Но этот путь трудозатратен, поэтому разработаны также автоматические меры оценки качества тонального анализа. Они не всегда точны для каждого конкретного текста, но признаны в качестве легитимных. Назовем только некоторые, наиболее популярные: *F*-мера гармонической точности (*F-measure of harmomic accuracy*), точность/достоверность (*precision*), полнота (*recall*).

Алгоритмические и программные решения для тонального анализа. Сегодня доступные научному сообществу программные инструменты включают более двух десятков алгоритмов; некоторые из них доработаны в готовые программные продукты для нескольких языков. Примером такого программного пакета может быть британский *SentiStrength* (Thelwall, 2013), применяемый для определения тональности и ее силы и сегодня работающий с испанским и другими европейскими языками. Его аналогом может выступать пакет *SocialMention*. Часто при оценке различных алгоритмов в один ряд ставятся алгоритмы и тезаурусы (например, *SentiStrength* и *SenticNet*). Это обусловлено размытыми границами между пони-

манием алгоритма и тезауруса, поскольку наиболее сложные тезаурусы де-факто устроены алгоритмически и являются не наборами размеченных лексем, а наборами лексем с приданной семантической связью, что позволяет словарю действовать как единое целое.

Более простые и доступные инструменты, которые использовались для решения задач майнинга мнений в середине 2000-х – начале 2010-х гг., включают такие известные инструменты, как *Profile of Mood States on Google* и *OpinionFinder* (Wilson et al., 2005), использовавшие простую бинарную (позитивные/негативные) классификацию мнений. Но сегодня предпочтительно использовать инструментарий, позволяющий решить большее количество задач и адаптированный для разных языков.

Тональный анализ за пределами англоязычной зоны: проблема отсутствия лексиконов и сентимент-анализ для русского языка

С начала 2010-х гг. двумя актуальными тенденциями в сентимент-анализе остаются его развитие для социальных медиа и развитие многоязычного (кросс-лингвального) тонального анализа.

Общим местом в научной литературе о сентимент-анализе стало указание на отсутствие исследований за пределами англоязычной зоны. Причиной этого чаще всего является отсутствие доступных тезаурусов для тонального анализа на национальных языках. До 2014 г. существовало всего 12 общедоступных неанглоязычных лексиконов для тонального анализа. В 2014 г. была предпринята беспрецедентная попытка создать словари для 136 языков и протестировать их эффективность на 30 языках, а также на 12 доступных неанглоязычных словарях (Chen, Skiena, 2014). Но степень сходимости результата с ручным кодированием была ниже 50%, и предложенные авторами словари практически не используются.

В 2016–2018 гг. появилось сразу много работ, в которых анализу подвергаются тексты на основных европейских языках, в т. ч. несколько кросс-лингвальных (см. ниже). Но в основном авторы не занимаются разработкой словарей для каждого из языков, а используют машинный перевод негативных и позитивных лексем из доступных англоязычных тезаурусов. Это, конечно, снижает качество анализа. Сегодня в исследованиях по всему миру используются два ведущих англоязычных семантических тезауруса:

- производные классического семантического словаря английского языка *WordNet – WordNet-Affect* и *SentiWordNet* 3.0 (Baccianella et al., 2010), разработанные Принстонским университетом;

- *SenticNet* 4 (http://sentic.net), разработанный *MIT*. Тезаурус содержит 50000 единиц и обладает несколькими преимуществами относительно других тезаурусов: он связывает слова семантически, т. е. более гибок, охватывает оценкой большее число случаев совстречаемости слов и может обеспечить связь *semantics* и *sentics* (когнитивной информации и аффективного модуса высказывания).

Русский язык пока редко становится объектом исследования в тональном анализе. Можно отметить работы Четверкина и Лукашевич (2012, 2013), в т. ч. сравнивающие подходы к тональному анализу, разработанные на тот момент. В 2011 г. Пазельская и Соловьев предложили метод выделения эмоциональной лексики в русском языке путем сочетания векторных и словарных методик (Пазельская, Соловьев, 2011), рассчитанный, однако, на «грамматически правильные тексты СМИ», а не на контент из Интернета. Зарубежные авторы, участвующие в разработке методик кросслингвального тонального анализа, иногда также используют русский язык как целевой (*target language),* то есть такой, на который совершается перевод словаря с английского языка.

Попытки создать собственные словари для русского языка предпринимались несколько раз. Так, молдавские исследователи (Sokolova, Bobicev, 2009) разработали первый лексикон для русского (и румынского) языка на основе *WordNet-Affect*, но данная разработка не применяется для изучения текстов в Интернете. Лаборатория Интернет-исследований НИУ ВШЭ завершает создание тонального словаря русского языка на основе краудсорсинга и экспертной разметки, но пока словарь недоступен. Некоторые словарные подходы и подходы на базе машинного обучения были опробованы участниками семинара РОМИП в начале 2010-х (Клековкина, Котельников, 2012); их опыт указывает на то, что для русского языка следует применять комбинированные методы сентимент-анализа.

В условиях отсутствия национального словаря, которому можно доверять, лексиконы, переведенные с английского, следует обогащать, в т. ч. добавлять в лексикон оцененный вручную частотный словарь самого кейса и использовать (после экспертной оценки) комплекты размеченной эмоциональной лексики из открытых источников (иных исследований, коммерческих словарей и т. д.).

Сентимент-анализ для Твиттера

Активное изучение твитов с точки зрения тонального анализа началось около 2010 г. (Kouloumpis et al., 2011). Изучение тональности текстов пользователей, особенно коротких текстов Твиттера, как показывают работы многих ученых, значительно осложнено в силу нескольких причин.

1) Выбор единицы анализа. Как указано выше, тональный анализ может проводиться на уровнях от леммы до всего документа. Но к Твиттеру возможно применение классификации как на уровне предложения, так и на уровне документа, что в данном случае не важно по сути (ведь твит крайне редко содержит более одного предложения), но меняет оптику исследователя, предлагаемые инструменты, результаты анализа и их интерпретацию. В воркшопе *SemEval* было предложено разделение семантического анализа для Твиттера на анализ всего сообщения (*message-level analysis*) и терминов (слов, словосочетаний, последовательностей слов; *term-level analysis*) (Poursepanj et al., 2013; Zhu et al., 2014).

2) Мультилингвальность твитов: часто хэштегами, написанными на одном языке (например, #*JeSuisCharlie*), маркируются твиты на другом языке. Анализ таких твитов требует многоязычных лексиконов и/или корпусов текстов для машинного обучения (Vilares et al., 2017) либо нахождения алгоритма фильтрации многоязычных твитов из корпусов текстов.

3) Языковые особенности твитов. Во-первых, это нарушение грамматических и иных языковых правил пользователями платформы. Во-вторых, отсутствие сложных предложений в силу лимита на число символов (140 знаков). В-третьих, большое количество жаргонной и просторечной лексики, аббревиатур и эмотиконов (смайликов) (Kontopoulos et al. 2013). На исключительном или частичном использовании эмотиконов для оценки тональности построено уже более десятка серьезных исследований тональности в Твиттере (Pak, Paroubek, 2010), но интерпретация эмотиконов в кросс-лингвальной среде пока мало изучена, несмотря на их кажущееся единообразие.

4) Высокий процент сарказма в корпусах твитов (Liu, 2012; Maynard, Greenwood, 2014). Такие твиты *запутывают* классификатор, так как позитивные лексемы в совокупности создают негативное высказывание за счет интонации и/или изменения порядка слов. В случае конфликтных дискуссий в Твиттере число таких твитов еще сильнее возрастает (Bodrunova et al., 2015). Для устра-

нения сбоев классификатора из-за сарказма рекомендуется применение методов машинного обучения на размеченных коллекциях твитов.

5) Особенности предварительной обработки наборов данных. Как правило, процедуры препроцессинга включают токенизацию (*tokenization*), нормализацию (*normalization*) и разметку по частям речи (*POS tagging*). Однако некоторые работы показали, что разметка по частям речи не работает для твитов (Kouloumpis et al., 2011).

В силу указанных причин инструменты, разработанные для корпусов текстов большего объема и большей ориентации на письменный стиль изложения, дают сниженный результат анализа (*low recall*) для корпусов твитов. Поэтому для Твиттера, во-первых, предложено использовать наиболее простые методы анализа, которые подвергаются постоянной ручной перепроверке (см. Nielsen, 2011), во-вторых, представлено несколько (пока недостаточно апробированных) многоступенчатых моделей тонального анализа на основе относительно простых шагов (Chikersal et al., 2015; Sudchev et al., 2014). Более сложные алгоритмы комбинируют выделение признаков изучаемого объекта (*feature-based models, tree kernel-based models* и др.) с *n*-граммами и словарями, а также с машинным обучением (Agarwal et al., 2011) или используют построение онтологий на базе формального анализа концептов (*formal concept analysis*) (Kontopoulos et al., 2013).

В целом тональный анализ для Твиттера пока находится на начальной стадии развития. Часто оценка сентимента твитов воспринимается как самоцель. Лишь в нескольких работах сентимент-анализ твитов применяется для анализа, предсказания или описания иных явлений. Так, сентимент твитов связывается с движением рынка акций (Smailovic et al., 2013), политическими предпочтениями граждан (Ceron et al., 2014; Wang et al., 2012; Zhou et al., 2013), осведомленностью во время техногенной катастрофы (Verma et al., 2011). При этом тональный анализ должен, в первую очередь, служить целям описания и предсказания социальных явлений и предупреждения разрушений, жертв и паники во время природных и техногенных катастроф.

Бросается в глаза отсутствие исследований, связывающих тональность речи пользователей и их статус в дискуссии (например, выдвижение на влиятельную позицию; Bodrunova et al., 2016) или же метаданные пользователя (регион проживания, социальный статус и т. д.) с его эмоциональной стратегией. Так, только в одной работе

(Stieglitz, Dang-Xuan, 2012) достоверно подтверждается связь тональности пользователей и их популярности по метрике *ретвиты*.

Аккумуляция опыта: современный многоязычный сентимент-анализ для социальных медиа

Кросс-лингвальный тональный анализ (*cross-language sentiment analysis, CLSA*) состоит в определении сентимента в нескольких параллельных корпусах текстов на разных языках при условии, что эталонная выборка размечена только для одного языка (Bader et al., 2011). С начала 2010-х гг. наблюдается растущий интерес к развитию методик *CLSA*. Но, как уже сказано выше, многоязычный тональный анализ находится в зависимости от качества машинного перевода и требует как минимум проверки независимости от перевода (*translator independency test*). Иногда применяется обратная стратегия перевода: переводятся не тезаурусы (с английского на целевые языки), а сами пользовательские тексты (с целевых языков на английский), после чего к ним применяются инструменты для работы с английским языком. При этом сравнение результатов тонального анализа текстов после машинного перевода почти не отличается от результатов анализа оригинальных неанглоязычных текстов (проверка проводилась более чем для полутора десятков алгоритмов; Araujo et al., 2016)); машинный перевод с использованием разных переводчиков также дает сходный результат.

В конце 2000-х было сделано много попыток улучшить качество кросс-лингвального тонального анализа путем применения комбинации тематического моделирования, методов синхронного и совместного машинного обучения и др. с тональным анализом на основе машинного перевода. Но в начале 2010-х вышел ряд работ (Duh et al., 2011; Balamurali et al., 2013), показавший, что в таких системах пока плохо учитывается разница в оформлении мнения и субъективности в разных языках. Для Твиттера, с его платформенными особенностями, кросс-лингвальный анализ становится затруднен еще более. Отдельные работы посвящены сентимент-анализу в переводах одного корпуса текстов на разные языки (Balahur et al., 2014; Becker et al., 2017); в таких работах часто используется метод триангуляции, включающий этапы перевода и ручной корректировки (Steinberger et al., 2011). Но для задач сопоставления реальных дискурсов в соцсетях этот метод неприменим. При этом очевидна необходимость учитывать не только

языковые, но и культурные и контекстуальные особенности изучаемых корпусов текстов при составлении словарей и разметке эталонных выборок (Bodrunova, 2018); следует считать такой анализ кросс-культурным тональным анализом.

Сегодня крайне редки работы, которые применяли бы сентимент-анализ на разных языках (включая русский) к текстам СМИ или социальных медиа для анализа сходных социальных конфликтов, острых вопросов повестки дня, войн, антропогенных и природных катастроф. Разработанные на данный момент программы и алгоритмы для кросс-лингвального тонального анализа в Твиттере, например *SentiSAIL* (Shalunts, Backfield, 2016) или *B4MSA* (Tellez et al., 2017), пока не применялись к анализу реальных тематически сходных выборок на разных языках.

В этом контексте важна работа Mozetic et al. (2016): авторы показывают, что при анализе твитов на разных языках следует идти простым, но трудозатратным путем и использовать труд кодировщиков для создания размеченных кейс-специфичных эталонных выборок. В этом случае для повышения качества тонального анализа критически значимым остается число закодированных для эталонной выборки твитов и степень согласия между экспертами, а не сам алгоритм сентимент-анализа. Исходя из авторских аргументов, для тонального анализа Твиттера должно применяться машинное обучение на основе кейс-специфичных экспертных словарей или выборок твитов. Подобный подход применялся в т. ч. для русскоязычных текстов (Yussupova et al., 2015).

Заключение: методика кросс-культурного тонального анализа для Твиттера

Вывод, который можно сделать из данного обзора, состоит в том, что пока не разработана методика тонального анализа на основе параллельных кейс-ориентированных лексиконов, сформированных вручную по одному и тому же принципу для сходных кейсов на разных языках (например, для дискуссий о межнациональных конфликтах, миграционном кризисе или природных катастрофах в разных языковых сегментах Твиттера).

Основой такой методики должна стать единая конструкция лексикона. Элементами такого лексикона для каждого из языков могли бы быть:

- *сердцевина* – узкая группа базовых тонально размеченных лексем данного языка;

- *расширение* для данного языка, пришедшее из единого источника путем машинного перевода (например, из тезауруса *WordNet* и его производных или тезауруса *SenticNet*);

- корпус культурно-ориентированной эмоциональной маркированной лексики, созданный путем экспертной очистки и тональной разметки частотного словаря данного кейса и сопоставленный с такими же корпусами для параллельных кейсов, с тем чтобы установить степень сходства;

- сходные процедуры проверки качества тонального анализа.

Такой дизайн исследования позволил бы оценить, насколько в принципе возможна автоматизация и получение сходных результатов тонального анализа без применения тотального перевода текстов на английский язык, что для Твиттера неприемлемо. Он также позволил бы разработать новые, универсальные критерии оценки качества тонального анализа коротких пользовательских текстов в сети Интернет.

Примечания

[1] Статья подготовлена при поддержке Гранта Президента РФ для молодых ученых – докторов наук МД-6259.2018.6 "Кросс-культурный тональный анализ пользовательских текстов в сети Интернет".

Библиография

Клековкина М. В., Котельников Е. В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // Труды XIV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». 2012. С. 118–123.

Пазельская А. Г., Соловьев А. Н.. Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: по мат. ежегодной Междунар. конф. «Диалог» (Бекасово, 25–29 мая 2011 г.). М.: Изд-во РГГУ, 2011.

Усталов Д. А. Извлечение терминов из русскоязычных текстов при помощи графовых моделей // Теория графов и приложения: мат. конф. М., 2012. С. 62–69.

Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau R. (2011) Sentiment Analysis of Twitter Data. In *Proc. ACL 2011 Workshop on Languages in Social Media*. Pp. 30–38.

Araujo M., Reis J., Pereira A., Benevenuto F. (2016) An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. Pp. 1140–1145.

Baccianella S., Esuli A., Sebastiani F. (2010) SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC Vol. 10*. Pp. 2200–2204.

Balahur A., Turchi M., Steinberger R., Ortega J. M. P., Jacquet G., Küçük D., El Ghali A. (2014) Resource Creation and Evaluation for Multilingual Sentiment Analysis in Social Media Texts. In *LREC*. Pp. 4265–4269.

Balamurali A. R., Khapra M. M., Bhattacharyya P. (2013) Lost in translation: viability of machine translation for cross language sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, Berlin, Heidelberg. Pp. 38–49.

Barbosa L., Feng J. (2010) Robust Sentiment Detection on Twitter from Biased and Noisy Data. In *Proc. of 23rd Int. Conf. on Computational Linguistics: Posters (COLING '10), Association for Computational Linguistics, Stroudsburg, PA, USA*. Pp. 36–44.

Becker K., Moreira V. P., dos Santos A. G. (2017) Multilingual emotion classification using supervised learning: Comparative experiments. *Information Processing & Management* 53 (3): 684–704.

Bodrunova S. S. (2018) When Context Matters. Analyzing Conflicts with the Use of Big Textual Corpora from Russian and International Social Media. *Partecipazione E Conflitto*11 (2): 497–510.

Bodrunova S. S., Blekanov I. S., Maksimov A. (2016) Measuring influencers in twitter ad-hoc discussions: active users vs. internal networks in the discourse on biryuliovo bashings in 2013. In *Artificial Intelligence and Natural Language Conference (AINL), IEEE*. Pp. 1–10.

Bodrunova S. S., Litvinenko A. A., Gavra D. P., Yakunin A. V. (2015) Twitter-Based Discourse on Migrants in Russia: The Case of 2013 Bashings in Biryulyovo. *International Review of Management and Marketing 5* (1S).

Ceron A., Curini L., Iacus S. M., Porro G. (2014) Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society* 16 (2): 340–358.

Chen, Y., Skiena, S. (2014) Building Sentiment Lexicons for All Major Languages. In *ACL (2)*. Pp. 383–389.

Chetviorkin A., Loukachevitch N. (2012) Extraction of Russian Sentiment Lexicon for Product Meta-Domain. In *Proc. of the 24th International Conference on Computational Linguistics (COLING), Bombay, India*. Pp. 593–610.

Chetviorkin A., Loukachevitch N. (2013) Evaluating Sentiment Analysis Systems in Russian. In *Proc. of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Sofia, Bulgaria*. Pp. 12–17.

Chikersal P., Poria S., Cambria, E. (2015) SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning. In *SemEval@ NAACL-HLT*. Pp. 647–651.

Cruz F. L., Troyano J. A., Pontes B., Ortega F. J. (2014) Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications* 41 (13): 5984–5994.

Dave K., Lawrence S., Pennock D. M. (2003) Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW*. Pp. 519–528.

Dos Santos C. N., Gatti, M. (2014) Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *COLING*. Pp. 69–78.

Duh K., Akinori F., Masaaki N. (2011) Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers (ACL-2011)*.

Gonçalves P., Araujo M., Benevenuto F., Cha M. (2013) Comparing and combining sentiment analysis methods. In *Proceedings of the 1 st ACM Conference on Online Social Networks (COSN), Boston, USA, ACM*. Pp. 27–38.

Heerschop B., Goossen F., Hogenboom A., Frasincar F., Kaymak U., de Jong F. (2011) Polarity analysis of texts using discourse structure. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. Pp. 1061–1070.

Kan D. (2012) *Rule-based approach to sentiment analysis at ROMIP 2011*. Available at: http://www. dialog-21. ru/digests/dialog2012/materials/pdf/Kan. pdf

Kontopoulos E., Berberidis C., Dergiades T., Bassiliades N. (2013) Ontology-based sentiment analysis of twitter posts. *Expert systems with applications* 40 (10): 4065–4074.

Kouloumpis E., Wilson T., Moore J. D. (2011) Twitter sentiment analysis: The good the bad and the omg! *Icwsm* 11 (538–541): 164.

Liu B. (2010) Sentiment Analysis and Subjectivity. In *Handbook of natural language processing*, *2*. Pp. 627–666.

Liu B. (2012) Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies 5* (1): 1–167.

Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E. (2015) SentiRuEval: testing object-oriented sentiment analysis systems in Russian. In *Proceedings of International Conference Dialog. Vol. 2*. Pp. 3–13.

Maynard D., Greenwood M. A. (2014) Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. In *LREC*. Pp. 4238–4243.

Mozetič I., Grčar M., Smailović J. (2016) Multilingual Twitter sentiment classification: The role of human annotators. *PloS one* 11 (5): e0155036.

Nakov P., Rosenthal S., Kiritchenko S., Mohammad S. M., Kozareva Z., Ritter A., Zhu X. (2016) Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. *Language Resources and Evaluation* 50 (1): 35–65.

Nielsen F. Å. (2011) A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *arXiv preprint arXiv:1103.2903*.

Ohmura M., Kakusho K., Okadome T. (2014) Social mood extraction from Twitter posts with document topic model. In *Information Science and Applications (ICISA), 2014 International Conference on*. Pp. 1–4.

Pak A., Paroubek P. (2010) Twitter as a corpus for sentiment analysis and opinion mining. In *LREc, Vol. 10, No. 2010*.

Pang B., Lee L. (2008) Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2 (1–2): 1–135.

Pang B., Lee L., Vaithyanathan S. (2002) Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*. Pp. 79–86.

Ponomareva N., Thelwall M. (2012) Do neighbours help?: an exploration of graph-based algorithms for cross-domain sentiment classification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Pp. 655–665.

Pontiki M., Galanis D., Papageorgiou H., Androutsopoulos I., Manandhar S., AL-Smadi M., Hoste V. (2016) SemEval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*. Pp.. 19–30.

Poursepanj H., Weissbock J., Inkpen D. (2013). uOttawa: System description for SemEval 2013 Task 2 Sentiment Analysis in Twitter. In *SemEval@ NAACL-HLT*. Pp. 380–383.

Rosenthal S., Farra N., Nakov P. (2017) SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Pp. 502–518.

Saif H., He Y., Alani H. (2012) Semantic sentiment analysis of twitter. *The Semantic Web–ISWC 2012*. Pp. 508–524.

Schuller B., Knaup T. (2011) Learning and knowledge-based sentiment analysis in movie review key excerpts. In *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*. Pp. 448–472.

Serrano-Guerrero J., Olivas J. A., Romero F. P., Herrera-Viedma E. (2015) Sentiment analysis: A review and comparative analysis of web services. *Information Sciences* 311: 18–38.

Shalunts G., Backfried G. (2016) Multilingual Sentiment Analysis on Data of the Refugee Crisis in Europe. In *DATA ANALYTICS 2016*. P. 57.

Sharma A., Dey S. (2013) A boosted SVM based sentiment analysis approach for online opinionated text. In *Proceedings of the 2013 Research in Adaptive and Convergent Systems*. Pp. 28–34.

Smailović J., Grčar M., Lavrač N., Žnidaršič M. (2013) Predictive sentiment analysis of tweets: A stock market application. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*. Pp. 77–88.

Sokolova M., Bobicev V. (2009) Classification of Emotion Words in Russian and Romanian Languages. In *RANLP*. Pp. 416–420.

Steinberger J., Lenkova P., Kabadjov M. A., Steinberger R., Van der Goot E. (2011) Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora. In *RANLP*. Pp. 770–775.

Stieglitz S., Dang-Xuan L. (2012) Political communication and influence through microblogging – An empirical analysis of sentiment in Twitter messages and retweet behavior. In *System Science (HICSS), 2012 45th Hawaii International Conference on*. Pp. 3500–3509.

Suchdev R., Kotkar P., Ravindran R., Swamy, S. (2014) Twitter Sentiment Analysis Using Machine Learning and Knowledge-Based Approach. *International Journal of Computer Applications* 103 (4).

Tellez E. S., Miranda-Jiménez S., Graff M., Moctezuma D., Suárez R. R., Siordia O. S. (2017) A Simple Approach to Multilingual Polarity Classification in Twitter. In *Pattern Recognition Letters*.

Thelwall M. (2013) Heart and soul: Sentiment strength detection in the social web with sentistrength. *Proceedings of the CyberEmotions* 5: 1–14.

Verma S., Vieweg S., Corvey W. J., Palen L., Martin J. H., Palmer M., Anderson K. M. (2011) Natural Language Processing to the Rescue? Extracting" Situational Awareness" Tweets During Mass Emergency. In *ICWSM*.

Vilares D., Alonso M. A., Gómez-Rodríguez C. (2017) Supervised sentiment analysis in multilingual environments. *Information Processing & Management* 53 (3): 595–607.

Wang H., Can D., Kazemzadeh A., Bar F., Narayanan S. (2012) A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*. Pp. 115–120.

Wilson T., Hoffmann P., Somasundaran S., Kessler J., Wiebe J., Choi Y., Cardie C., Riloff E., Patwardhan S. (2005) OpinionFinder: A system for subjectivity analysis. In *Proceedings HLT/EMNLP, Vancouver (BC)*. Pp. 34–35.

Yessenov K., Misailovic S. (2009) Sentiment analysis of movie review comments. *Methodology* 17: 1–7.

Yussupova N., Bogdanova D., Boyko M. (2012) Applying of sentiment analysis for texts in Russian based on machine learning approach. In *Proceedings of Second International Conference on Advances in Information Mining and Management*. Pp. 8–14.

Zhou X., Tao X., Yong J., Yang Z. (2013) Sentiment analysis on tweets for social events. In *Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on*. Pp. 557–562.

Zhu X., Kiritchenko S., Mohammad S. (2014) NRC-Canada-2014: Recent Improvements in the Sentiment Analysis of Tweets. In *SemEval@ COLING*. Pp. 443–447.

References

Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau R. (2011) Sentiment Analysis of Twitter Data. In *Proc. ACL 2011 Workshop on Languages in Social Media*. Pp. 30–38.

Araujo M., Reis J., Pereira A., Benevenuto F. (2016) An Evaluation of Machine Translation for Multilingual Sentence-Level Sentiment Analysis. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. Pp. 1140–1145.

Baccianella S., Esuli A., Sebastiani F. (2010) SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC Vol. 10*. Pp. 2200–2204.

Balahur A., Turchi M., Steinberger R., Ortega J. M. P., Jacquet G., Küçük D., El Ghali A. (2014) Resource Creation and Evaluation for Multilingual Sentiment Analysis in Social Media Texts. In *LREC*. Pp. 4265–4269.

Balamurali A. R., Khapra M. M., Bhattacharyya P. (2013) Lost in Translation: Viability of Machine Translation for Cross-Language Sentiment Analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, Berlin, Heidelberg. Pp. 38–49.

Barbosa L., Feng J. (2010) Robust Sentiment Detection on Twitter from Biased and Noisy Data. In *Proc. of 23rd Int. Conf. on Computational Linguistics: Posters (COLING '10), Association for Computational Linguistics, Stroudsburg, PA, USA*. Pp. 36–44.

Becker K., Moreira V. P., dos Santos A. G. (2017) Multilingual Emotion Classification Using Supervised Learning: Comparative Experiments. *Information Processing & Management* 53 (3): 684–704.

Bodrunova S. S. (2018) When Context Matters. Analyzing Conflicts with the Use of Big Textual Corpora from Russian and International Social Media. *Partecipazione E Conflitto*11 (2): 497–510.

Bodrunova S. S., Blekanov I. S., Maksimov A. (2016) Measuring Influencers in Twitter Ad-Hoc Discussions: Active Users vs. Internal Networks in the Discourse on Biryuliovo Bashings in 2013. In *Artificial Intelligence and Natural Language Conference (AINL), IEEE*. Pp. 1–10.

Bodrunova S. S., Litvinenko A. A., Gavra D. P., Yakunin A. V. (2015) Twitter-Based Discourse on Migrants in Russia: The Case of 2013 Bashings in Biryulyovo. *International Review of Management and Marketing 5* (1S).

Ceron A., Curini L., Iacus S. M., Porro G. (2014) Every Tweet Counts? How Sentiment Analysis of Social Media Can Improve our Knowledge of Citizens' Political Preferences with an Application to Italy and France. *New Media & Society* 16 (2): 340–358.

Chen, Y., Skiena, S. (2014) Building Sentiment Lexicons for All Major Languages. In *ACL (2)*. Pp. 383–389.

Chetviorkin A., Loukachevitch N. (2012) Extraction of Russian Sentiment Lexicon for Product Meta-Domain. In *Proc. of the 24th International Conference on Computational Linguistics (COLING), Bombay, India*. Pp. 593–610.

Chetviorkin A., Loukachevitch N. (2013) Evaluating Sentiment Analysis Systems in Russian. In *Proc. of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Sofia, Bulgaria*. Pp. 12–17.

Chikersal P., Poria S., Cambria, E. (2015) SeNTU: Sentiment Analysis of Tweets by Combining a Rule-Based Classifier with Supervised Learning. In *SemEval@ NAACL-HLT*. Pp. 647–651.

Cruz F. L., Troyano J. A., Pontes B., Ortega F. J. (2014) Building Layered, Multilingual Sentiment Lexicons at Synset and Lemma Levels. *Expert Systems with Applications* 41 (13): 5984–5994.

Dave K., Lawrence S., Pennock D.M. (2003) Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proceedings of WWW*. Pp. 519–528.

Dos Santos C. N., Gatti, M. (2014) Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *COLING*. Pp. 69–78.

Duh K., Akinori F., Masaaki N. (2011) Is Machine Translation Ripe for Cross-Lingual Sentiment Classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers (ACL-2011)*.

Gonçalves P., Araujo M., Benevenuto F., Cha M. (2013) Comparing and Combining Sentiment Analysis Methods. In *Proceedings of the 1 st ACM Conference on Online Social Networks (COSN), Boston, USA, ACM*. Pp. 27–38.

Heerschop B., Goossen F., Hogenboom A., Frasincar F., Kaymak U., de Jong F. (2011) Polarity Analysis of Texts Using Discourse Structure. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. Pp. 1061–1070.

Kan D. (2012) *Rule-Based Approach to Sentiment Analysis at ROMIP 2011*. Available at: http://www. dialog-21. ru/digests/dialog2012/materials/pdf/Kan. pdf

Klekovkina M. V., Kotel'nikov E. V. (2012) Metod avtomaticheskoy klassifikatsii tekstov po tonal'nosti, osnovannyy na slovare emotsional'noy leksiki [Method of Automatic Text Classification by Sentiment Based on the Dictionary of Sentiment Lexicon]. In *Trudy XIV Vserossiyskoy nauchnoy konferentsii «Elektronnyye biblioteki: perspektivnyye metody i tekhnologii, elektronnyye kollektsii* [Proceedings of the 14th All-Russian scientific conference "Electronic Libraries: Promising methods and Technologies, Electronic Collections]. Pp. 118–123. (In Russian)

Kontopoulos E., Berberidis C., Dergiades T., Bassiliades N. (2013) Ontology-Based Sentiment Analysis of Twitter Posts. *Expert systems with applications* 40(10): 4065–4074.

Kouloumpis E., Wilson T., Moore J. D. (2011) Twitter Sentiment Analysis: the Good the Bad and the Omg! *Icwsm* 11(538–541): 164.

Liu B. (2010) Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing*, *2*. Pp. 627– 666.

Liu B. (2012) Sentiment Analysis and Opinion Mining. In *Synthesis lectures on human language technologies 5* (1): 1–167.

Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E. (2015) SentiRuEval: Testing Object-Oriented Sentiment Analysis Systems in Russian. In *Proceedings of International Conference Dialog. Vol. 2*. Pp. 3–13.

Maynard D., Greenwood M. A. (2014) Who Cares About Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. In *LREC*. Pp. 4238–4243.

Mozetič I., Grčar M., Smailović J. (2016) Multilingual Twitter Sentiment Classification: the Role of Human Annotators. *PloS one* 11 (5): e0155036.

Nakov P., Rosenthal S., Kiritchenko S., Mohammad S. M., Kozareva Z., Ritter A., Zhu X. (2016) Developing a Successful SemEval Task in Sentiment Analysis of Twitter and Other Social Media Texts. *Language Resources and Evaluation* 50 (1): 35–65.

Nielsen F. Å. (2011) A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In *arXiv preprint arXiv:1103.2903*.

Ohmura M., Kakusho K., Okadome T. (2014) Social Mood Extraction from Twitter Posts with Document Topic Model. In *Information Science and Applications (ICISA), 2014 International Conference on*. Pp. 1–4.

Pak A., Paroubek P. (2010) Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREc, Vol. 10, No. 2010*.

Pang B., Lee L. (2008) Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval* 2 (1–2): 1–135.

Pang B., Lee L., Vaithyanathan S. (2002) Thumbs up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*. Pp. 79–86.

Pazel'skaya A. G., Solov'yev A. N. (2011) Metod opredeleniya emotsiy v tekstakh na russkom yazyke [The Method of Sentiment Analysis of Texts in Russian] In *Komp'yuternaya lingvistika i intellektual'nyye tekhnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii «Dialog» (Bekasovo, 25–29 maya 2011 g.)* [Computer Linguistics and Intelligence Technologies: based on the proceedings of the annual International Conference "Dialog" (Bekasovo, May 25–29, 2011]. Moscow: RGGU Publ. (In Russian)

Ponomareva N., Thelwall M. (2012) Do Neighbours Help?: an Exploration of Graph-Based Algorithms for Cross-Domain Sentiment Classification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Pp. 655–665.

Pontiki M., Galanis D., Papageorgiou H., Androutsopoulos I., Manandhar S., AL-Smadi M., Hoste V. (2016) SemEval-2016 task 5: Aspect-Based Sentiment Analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*. Pp.. 19–30.

Poursepanj H., Weissbock J., Inkpen D. (2013) uOttawa: System Description for SemEval 2013 Task 2 Sentiment Analysis in Twitter. In *SemEval@ NAACL-HLT*. Pp. 380–383.

Rosenthal S., Farra N., Nakov P. (2017) SemEval-2017 task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Pp. 502–518.

Saif H., He Y., Alani H. (2012) Semantic Sentiment Analysis of Twitter. *The Semantic Web–ISWC 2012*. Pp. 508–524.

Schuller B., Knaup T. (2011) Learning and Knowledge-Based Sentiment Analysis in Movie Review Key Excerpts. In *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*. Pp. 448–472.

Serrano-Guerrero J., Olivas J. A., Romero F. P., Herrera-Viedma E. (2015) Sentiment Analysis: a Review and Comparative Analysis of Web Services. *Information Sciences* 311: 18–38.

Shalunts G., Backfried G. (2016) Multilingual Sentiment Analysis on Data of the Refugee Crisis in Europe. In *DATA ANALYTICS 2016*. P. 57.

Sharma A., Dey S. (2013) A Boosted SVM Based Sentiment Analysis Approach for Online Opinionated Text. In *Proceedings of the 2013 Research in Adaptive and Convergent Systems*. Pp. 28–34.

Smailović J., Grčar M., Lavrač N., Žnidaršič M. (2013) Predictive Sentiment Analysis of Tweets: A Stock Market Application. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*. Pp. 77–88.

Sokolova M., Bobicev V. (2009) Classification of Emotion Words in Russian and Romanian Languages. In *RANLP*. Pp. 416–420.

Steinberger J., Lenkova P., Kabadjov M. A., Steinberger R., Van der Goot E. (2011) Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora. In *RANLP*. Pp. 770–775.

Stieglitz S., Dang-Xuan L. (2012) Political Communication and Influence Through Microblogging – An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior. In *System Science (HICSS), 2012 45th Hawaii International Conference on*. Pp. 3500–3509.

Suchdev R., Kotkar P., Ravindran R., Swamy, S. (2014) Twitter Sentiment Analysis Using Machine Learning and Knowledge-Based Approach. *International Journal of Computer Applications* 103 (4).

Tellez E. S., Miranda-Jiménez S., Graff M., Moctezuma D., Suárez R. R., Siordia O. S. (2017) A Simple Approach to Multilingual Polarity Classification in Twitter. In *Pattern Recognition Letters*.

Thelwall M. (2013) Heart and Soul: Sentiment Strength Detection in the Social Web with Sentistrength. *Proceedings of the CyberEmotions* 5: 1–14.

Ustalov D. A. (2012) Izvlecheniye terminov iz russkoyazychnykh tekstov pri pomoshchi grafovykh modeley [Deriving Terms from Russian Texts by Means of Graph Models]. In *Teoriya grafov i prilozheniya. Materialy konferentsii* [Graphs Theory and Applications. Proceedings of the conference]. Moscow. Pp. 62–69. (In Russian)

Verma S., Vieweg S., Corvey W. J., Palen L., Martin J. H., Palmer M., Anderson K. M. (2011) Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency. In *ICWSM*.

Vilares D., Alonso M. A., Gómez-Rodríguez C. (2017) Supervised Sentiment Analysis in Multilingual Environments. *Information Processing & Management* 53 (3): 595–607.

Wang H., Can D., Kazemzadeh A., Bar F., Narayanan S. (2012) A System for Real-Time Twitter Sentiment Analysis of 2012 US Presidential Election Cycle. In *Proceedings of the ACL 2012 System Demonstrations*. Pp. 115–120.

Wilson T., Hoffmann P., Somasundaran S., Kessler J., Wiebe J., Choi Y., Cardie C., Riloff E., Patwardhan S. (2005) OpinionFinder: A System for Subjectivity Analysis. In *Proceedings HLT/EMNLP, Vancouver (BC)*. Pp. 34–35.

Yessenov K., Misailovic S. (2009) Sentiment Analysis of Movie Review Comments. *Methodology* 17: 1–7.

Yussupova N., Bogdanova D., Boyko M. (2012) Application of Sentiment Analysis to Texts in Russian Based on the Machine Learning Approach. In *Proceedings of Second International Conference on Advances in Information Mining and Management*. Pp. 8–14.

Zhou X., Tao X., Yong J., Yang Z. (2013) Sentiment Analysis on Tweets for Social Events. In *Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on*. Pp. 557–562.

Zhu X., Kiritchenko S., Mohammad S. (2014) NRC-Canada-2014: Recent Improvements in the Sentiment Analysis of Tweets. In *SemEval@ COLING*. Pp. 443–447.

Поступила в редакцию
01.10.2018