УДК 81’42

Зверева П.Л.
Московский государственный областной университет

СЕНТИМЕНТ-АНАЛИЗ ТЕКСТА (НА МАТЕРИАЛЕ ПЕЧАТНЫХ ТЕКСТОВ ГАЗЕТЫ «THE NEW YORK TIMES» О РОССИИ И РОССИЯНАХ)

Аннотация. В статье исследуется эмоциональная оценка текста, в частности эмоциональная оценка текстов средств массовой информации. Рассматриваются такие понятия, как медиатекст, медиалингвистика, тональность (сентимент) текста. Проводится сентимент-анализ фрагментов печатных статей одного из ведущих изданий США, извлечённых из корпуса методом текстологического анализа и по ключевым словам. Полученные в результате сентимент-анализа данные сравниваются с результатами анкетного опроса, проведённого среди группы респондентов.
Ключевые слова: тональность, сентимент-анализ, медиатекст, медиалингвистика, средства массовой информации.

P. Zvereva
Moscow State Regional University

SENTIMENT-ANALYSIS OF TEXT (TEXTS ABOUT RUSSIA AND THE RUSSIANS FROM THE NEW YORK TIMES)

Abstract. The research of text affective evaluation, particularly the affective evaluation of Mass Media texts is described in the article. The concepts of media text, medialinguistics, and text sentiment are reviewed. The paper offers a sentiment-analysis of one of the US leading newspapers text fragments which were extracted from text corpus by means of textual analysis and key words. Received data are compared with the results of survey that was conducted among Russian-speaking respondents who can speak English.
Key words: sentiment, sentiment analysis, media text, medialinguistics, Mass Media.

В современном обществе человек узнаёт о происходящих в мире событиях, в основном, благодаря средствам массовой информации. СМИ оказывают влияние на способ мышления человека, его мировоззрение, поэтому с ними непосредственно связана проблема воздействия языка на мышление и поведение человека. Сегодня СМИ можно назвать источником и проводником коллективного знания, при этом они никогда не остаются индифферентными по отношению к тому, что освещают.

Изучением текстов СМИ занимаются такие сравнительно новые направления языкознания, как медиалингвистика (изучает языковые явления, сочетающие в себе разные семиотические коды, характерные для функционирования языка в сфере массовой коммуникации [1, с. 27–32]) и имагология – дисциплина, занимающаяся изучением образов «других», «чужих» наций, стран, культур, ино-

родных для воспринимающего субъекта [4, с. 31].

Известно, что СМИ в целом и медиатексты в частности обладают способностью влиять на массовое сознание реципиентов, создавать в их сознании определённые образы объектов, людей и даже стран [2, с. 29]. Сегодня авторы и редакция различных печатных изданий выступают в роли манипуляторов, формирующих общественное мнение, создавая тексты, в которых они «нужным» для себя способом освещают мировые события и создают «нужную» им картину мира. В результате у реципиентов складывается определённое представление о той или иной стране, её руководителях и жителях. Тем самым, образ страны продуктивно создаётся языковыми средствами СМИ. Авторы печатных текстов активно прибегают к использованию различных приёмов с целью создания нужной картины того или иного события в сознании реципиента, в нашем случае – читателя.

В этом контексте большой интерес представляет изучение тональности текстов СМИ. *Тональность текста* – это эмоциональная оценка, выраженная в данном тексте [3, с. 511]. Анализ тональности текста позволяет извлекать из текста эмоционально окрашенную лексику и получать представление об эмоциональном отношении авторов к объектам, о которых идёт речь в тексте, что, по нашему мнению, влияет на формирование определённого образа у реципиентов.

Анализ тональности текста (*сентимент-анализ*, англ. *sentiment analysis*) – это область компьютерной лингвистики, которая занимается изучением мнений и эмоций в текстовых документах и представляет собой совокупность методов, рассчитанных на автоматическое выявление эмоциональной оценки (*тональности*, или *сентимента*), выраженной в тексте. Такой анализ позволяет охарактеризовать эмоциональную окраску текста – положительную, отрицательную или нейтральную, выявить субъект и объект этого текста. Сентимент-анализ находит практическое применение в различных областях знаний (социологии, политологии, маркетинге и пр.)

Существует значительное число систем автоматической оценки эмоциональной окраски текста, базирующихся на различных методах (метод, основанный на правилах; метод, основанный на словарях; машинное обучение с учителем; машинное обучение без учителя; гибридный метод). Большинство систем являются продуктами коммерческих организаций, без свободного к ним доступа; ряд систем имеет свободный доступ, что позволяет проводить сентимент-анализ текстовых массивов. Одна из таких программ – система сентимент-анализа *Semantria* [6]. Система представляет собой расширение для программы Microsoft Excel и позволяет классифицировать тональность сообщений на нескольких европейских языках. Общая тональность текста определяется как положительная, отрицательная и нейтральная, по шкале от -1,5 до +1,5. Особенности системы: определение языка сообщения, создание пользовательских категорий объекта тональности, визуализация полученных данных.

Для нашего исследования был проведён сентимент-анализ печатных текстов, опубликованных в интернет-изда-

нии газеты «The New York Times» за 2013 г., и выявлена их тональность. Выбор данного издания не был случайным: газета «The New York Times» – ежедневная широкоформатная газета США, входящая в десятку наиболее популярных газет страны. Газета основана в 1851 г., и, как большинство американских газет, «The New York Times» была создана как региональное издание.

Анализируемый корпус текста представляет собой выборку из статей, сформированную по следующим критериям:

1) фрагмент текста должен содержать ключевые слова «Russia», «Russian», «Moscow», «Putin», «Soviet»;

2) методом первичного текстологического анализа определялись фрагменты, содержащие не только описательную характеристику, а также оценку сообщения.

В результате было отобрано 10 фрагментов статей о России и россиянах общим объёмом 525 слов, 3240 знаков. Сентимент-анализ показал следующие результаты: 6 фрагментов программа оценила как имеющие отрицательную тональность, 3 – нейтральную, 1 – положительную. Приведём примеры фрагментов, соответствующие каждой характеристике (см. табл.).

Таблица

Фрагменты текстов газетных статей о России и россиянах разной тональности

Исходный текст	Тональность документа	
	Тональность	Эмоциональные метки
The final straw appeared to be a law signed by Mr. Putin in December prohibiting the adoption of Russian children by American citizens, which the Kremlin said was retaliation for a new American law punishing Russian human rights violators. ("Another Reset With Russia in Obama's Second Term", February 1, 2013) [5]	-0,67499995	Отрицательная/ negative
Mr. Putin and other supporters of the Games have made clear that they view the pride and prestige of hosting the Olympics to be priceless. It is Russia's first Winter Olympics and its first Olympics since the Summer Games of 1980 in Moscow, when the United States led a boycott to protest the Soviet Union's invasion of Afghanistan. ("Putin's Vision of Olympic Glory Meets a More Earthbound Reality in Sochi", February 6, 2013) [5]	0,0128086	Нейтральная/ neutral
Roshen was doing so well in Russia partly because it introduced a Russian Classic line of chocolates, reviving 18 Soviet brands like the Seagull bar, a plain milk chocolate slab with a Socialist Realist style beach scene on the wrapper. ("Chocolate Factory, Trade War Victim", October 29, 2013) [5]	0,26970103	Положительная/ positive

Затем был проведён анкетный опрос среди русскоязычных респондентов, владеющих английским языком и выступающих в качестве экспертов, и проверена достоверность полученных результатов и эффективность работы системы сентимент-анализа *Semantria*. Этим респондентам было предложено определить эмоциональную оценку сообщения (возможные варианты: «положительная», «нейтральная», «отрицательная»).

В опросе приняли участие 10 респондентов. Отметим неоднозначный характер толкования тональности респондентами, так как только по отношению к 2 фрагментам из 10 оценки всех респондентов совпали. Это указывает на спорность определения тональности текста не только системами автоматической оценки эмоциональной окраски текста, но и респондентами. Мнения респондентов об этих фрагментах совпало с результатами, полученными в результате обработки текста системой сентимент-анализа *Semantria*:

Environmentalists have cited illegal dumping, destruction of forests and wildlife, and other violations. Dozens of residents say they have been forcibly relocated from their homes without adequate compensation, while thousands of others accepted payments and agreed to move to make way for construction. While local criticism of the Olympics is hardly unique to Sochi, some complaints have been addressed with classic Russian heavy-handedness. On Wednesday evening, a half-hour before a scheduled news conference about property disputes, organizers of the event were ejected from the hotel conference room they had booked and forced to gather on the street. ("Putin's Vision of Olympic Glory Meets a More Earthbound Reality in Sochi", February 6, 2013 [5]. Semantria: Тональность = -0,38456478714942 «отрицательная».)

Because American officials do not want to worsen the relationship and still hope for cooperation, they declined to publicly describe the plans. But within the administration it is taken for granted that the relationship with Russia is far less of a priority. ("Another Reset with Russia in Obama's Second Term", February 1, 2013 [5]. Semantria: Тональность = 0,013475000858306 «нейтральная».)

Well, as secretary of state today you get to deal with Vladimir Putin, who was born on third base and thinks he hit a triple. That is, even though Russia's economy is hugely corrupt and nowhere nearly as innovative as it should be, Putin sits atop a huge reserve of oil and gas that makes him think he's a genius and doesn't need to listen to anyone. ("Break All the Rules", January 22, 2013 [5]. Semantria: Тональность = -0,11430607736110 «отрицательная».)

В ходе нашего опроса выявлено, что 7 и более респондентов оценили тональность пяти фрагментов текста так же, как и система сентимент-анализа. При этом мнение о тональности некоторых фрагментов существенно колебалось. Так, 2 респондента присвоили приведённому ниже примеру «положительную» эмоциональную метку, 7 респондентов – «нейтральную» и 1 респондент – «отрицательную»:

Sochi, known as the capital of the Russian Riviera with its palm-lined promenades, beaches and Soviet-era sanitariums, is the first subtropical host of the Winter Games. That is remotely plausible only because the city sits at the foot of the Caucasus Mountains, the site of most of the outdoor events. ("Putin's Vision of Olympic

Glory Meets a More Earthbound Reality in Sochi", February 6, 2013 [5]. Semantria: Тональность = 0 «нейтральная».)

Респонденты и система сентимент-анализа *Semantria* дали разные оценки эмоциональной окраски текста по двум фрагментам:

While Russian lawmakers debated a bill that would outlaw "homosexual propaganda," nationalist and religious demonstrators on Friday attacked gay rights advocates who had gathered outside the lower house of Parliament to protest the legislation. ("Propaganda' by Gays Faces Russian Curbs Amid Unrest", January 25, 2013 [5]. Semantria: Тональность = -0,40000000596046 «отрицательная».)

Roshen was doing so well in Russia partly because it introduced a Russian Classic line of chocolates, reviving 18 Soviet brands like the Seagull bar, a plain milk chocolate slab with a Socialist Realist style beach scene on the wrapper. But this year, Roshen has missed Teacher's Day in Russia, a big day for giving chocolate gift boxes. ("Chocolate Factory, Trade War Victim", October 29, 2013 [5]. Semantria: Тональность = 0,26970103383064 «положительная».)

Отметим, что по отношению к первому фрагменту мнения респондентов разделились практически поровну (60% – «нейтральная» оценка, 40% – «отрицательная», оценка системы сентимент-анализа *Semantria* – «отрицательная»), и можно предположить, что оценки системы сентимент-анализа *Semantria* и респондентов совпали.

Опрос респондентов о втором фрагменте, приведённом выше, показал следующие результаты: «положительная» оценка – 0%, «нейтральная» – 90%, «отрицательная» – 10%. Здесь эмоциональная оценка респондентов не совпадает с «положительной» оценкой системы сентимент-анализа *Semantria*. Ни один респондент не оценил положительно окраску этого фрагмента статьи, что свидетельствует о полном несоответствии классификации тональности. Следовательно, можно утверждать, что интерпретация тональности фрагментов системой сентимент-анализа *Semantria* неоднозначна.

В ходе сентимент-анализа выявлено, что фрагментов с отрицательной тональностью больше, чем фрагментов с положительной тональностью, а результаты опроса респондентов показали, что наибольшее количество фрагментов имеет «нейтральную» оценку тональности. Мы предполагаем, что при определении оценки тональности респондент рассматривает фрагмент текста целиком, в то время как система сентимент-анализа анализирует отдельные слова и словосочетания. Отметим также, что система сентимент-анализа неверно интерпретирует в текстах иронию, сарказм, объективные предложения с эмоционально окрашенной лексикой.

В заключение отметим, что повышение точности – центральная задача автоматической оценки эмоциональной окраски текста. Для верного определения тональности сообщений необходима разработка метода, учитывающего не только синтаксические и морфологические, но и семантические особенности текста. Кроме того, необходимо создавать и пополнять так называемые тональные словари, представляющие собой списки слов со значением тональности для каждого отдельного слова.

Анализ тональности текста представляет значительный интерес для исследования печатных текстов СМИ

о России, так как он способствует уточнению точки зрения автора и выявлению его отношения к описываемым событиям и лицам. Эти знания помогают выявить приёмы, с помощью которых авторы создают нужный им смысл и интонацию повествования, что влияет на восприятие заложенной в тексте информации реципиентом.

Таким образом, сентимент-анализ можно рассматривать как ещё один формализованный метод, применяемый для изучения формирования образа страны и выявления приёмов манипуляции сознанием реципиента.

В результате проведённого исследования были выявлены проблемы, с которыми сталкиваются современные системы сентимент-анализа текста. Так, точность работы напрямую зависит от автоматического исключения из анализируемой информации нерелевантных сообщений, верного интерпретирования объективных предложений, а также решения проблемы анализа вопросительных и условных предложений, отрицания и др. Путём снятия этих проблем можно добиться построения эффективной системы сентимент-анализа, что позволит давать заключение о предвзятости выраженного мнения, о влиятельности авторов сообщений в СМИ и представлять результат работы в простой форме, доступной к использованию неспециалистами.

ЛИТЕРАТУРА:

1. Добросклонская Т.Г. Вопросы изучения медиатекстов. Опыт исследования современной английской медиаречи. М.: УРСС Эдиториал, 2005. 288 с.
2. Максименко О.И., Зверева П.П. Современные направления лингвистических исследований имиджа страны и её жителей // Вестник Московского государственного областного университета. Серия: Лингвистика. 2013. № 6. С. 25–30.
3. Пазельская А.Г., Соловьев А.Н. Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2011»: конференция. М., 2011. С. 510–522.
4. Папилова Е.В. Имагология как гуманитарная дисциплина // Вестник МГГУ им. М.А. Шолохова. Филологические науки. 2011. № 4. С. 31–40.
5. The New York Times Breaking News, World News and Multimedia. [Электронный ресурс]. URL: http://www.nytimes.com/ (дата обращения: 15.06.2014).
6. Semantria for Excel. [Электронный ресурс]. URL: https://semantria.com/excel (дата обращения: 16.06.2014).