УДК 811.161.1:004.4

АНАЛИЗ НОВОСТНЫХ СООБЩЕНИЙ САЙТА МИД РФ МЕТОДОМ СЕНТИМЕНТ-АНАЛИЗА

(статья 2)

М.В. Беляков

Московский государственный институт международных отношений
(университет) МИД России
пр-т Вернадского, 76, Москва, Россия, 119454
belmax007@hotmail.com

Оценка тональности текста представляет собой одну из самых актуальных и современных задач прикладной лингвистики. Сентимент-анализ, являясь одним из формальных методов анализа текста, позволяет, с одной стороны, оценивать эмоциональную окрашенность текста, а с другой — делать содержательные выводы по результатам выявленного сентимента. В период информационных войн эта задача относится к задачам первостепенной важности. В статье рассматривается вариант сентмент-анализа на примере ряда категорий, сформированных по материалам текстов сайта МИД РФ за один месяц. Для проведения анализа были получены 2 словаря — положительно и отрицательно маркированных основ слов. Результаты обработки текстов программой сентимент-анализа позволили выделить и графически представить оценку тональности текстов выделенных категорий, что подтвердило возможность использования данного формализованного метода для оценки и формирования общественного мнения.

Ключевые слова: дипломатический дискурс, сентимент-анализ, текст, сайт, МИД РФ

ВВЕДЕНИЕ

Автоматический анализ тональности (сентимент-анализ), выделяясь среди прочих современных методов формального анализа текстовой информации, декларирует в качестве своей основной задачи разработку алгоритмов по автоматическому структурированию текстовых данных и определение отзывов о них, т.е. объектом изучения сентимент-анализа являются выраженные в текстовой форме эмоции, мнения и оценки людей о каких-либо сущностях. Сентимент-анализ называют в зарубежной литературе также “opinion mining”, “sentiment extraction”, “subjectivity analysis” и др. К числу объектов, мнения людей о которых могут быть релевантными для подобного изучения, относятся продукты потребления, их атрибуты, службы по предоставлению различного рода услуг, организации, отдельные личности, события в мире, разнообразные темы, обсуждаемые в мировом сообществе, и т.д.

Данная область науки стала активно развиваться относительно недавно, и хотя первые исследования начались еще в 70—80-е гг. XX в. в работах таких исследователей, как Jaime Carbonell, Yorick Wilks и Janusz Bien, наибольшее число работ в этом направлении относится к концу 1990-х — началу 2000-х гг. К самым ранним из них относятся работы Hatzivassiloglou и McKeown (1997); Hearst (1992); Wiebe (1990—1994); Bruce и O'Hara (1999). Начиная с 2001 г. анализ тональности

стал одним из наиболее активно исследуемых разделов компьютерной лингвистики, в результате чего в последующие годы появилось значительное число работ по сентимент-анализу и извлечению мнений. Причиной столь пристального внимания исследователей к данной проблеме послужило сразу несколько факторов. Самым главным из них являлся активный рост сети Интернет, что открыло большие возможности для исследования эмоциональной составляющей в тексте, поскольку исследователям стало доступно бесконечное количество записанных в цифровой форме данных, выражающих в той или иной форме мнения людей. Кроме того, анализ тональности предоставил новые возможности для бизнеса, стал важной частью социологических исследований, нашел применение в коммерческой и промышленной сферах и впоследствии распространился на многие другие сферы общественной жизни. Все это послужило стимулом для активных исследований в данной области. Наконец, анализ тональности сам по себе открывал перед исследователями обширное поле для изучения, ставя перед ними множество подзадач, решение которых, в свою очередь, могло принести вклад в другие разделы исследований по обработке текстов на естественных языках.

Материалом для данного исследования послужили статьи сайта Министерства иностранных дел Российской Федерации. Статьи были опубликованы на сайте в разделе «Новости» с 1 по 28 февраля 2015 г. В выборку вошли все статьи, занимающие, по результатам контент-анализа [Беляков 20016], большее пространство в вербальном контенте поликодового сайта по сравнению с остальными. За единицу анализа был принят отрезок текста, соответствующий одной из следующих категорий: 1) украинский вопрос; 2) сотрудничество России с Китаем; 3) отношения России и Украины; 4) конфликт в Сирии; 5) сотрудничество с Туркменистаном; 6) отношения России и Греции; 7) санкции против России; 8) дипломатия сегодня.

ИНСТРУМЕНТ СЕНТИМЕНТ-АНАЛИЗА

Для проведения сентимент-анализа было составлено два словаря, которые включали в себя 300 основ слов положительно окрашенной лексики 390 основ отрицательно окрашенных слов. Программа производит поиск элементов из словаря в тексте, и в случае обнаружения слову приписывается соответствующая оценка «+1» или «–1». После того как программа проанализировала весь текст, оценки суммируются и выдается результат — сумма этих чисел.

Словарь положительно окрашенной лексики включал в себя следующие элементы: *автоном, авторитет, адаптац, адапти, адвокат, адекват, альтруист, ангел, аплод, баланс, безболез, безопас, безупреч, бесплатн, благо, благодар, благополуч, благоприят, благосл, блажен, блеск, блест, близ, блистательн, богат, бодр, божествен, братск, вдохнов, вежлив, великолеп, верн, весел, взбадр, вклад, внима, внимат, возмож, возрожд, восторг, восторж, восхитит, восхищ, выгод, выдающ, выигр, гарант, гостеприимн, действен, динамич, дипломат, добр, довер, доволен, довольн, достиж, достижен, достоин, достойн, доступ, друж, желае, желан, желат, желаю, забав, заветн, закон, заманчив, замет, замечательн, заслуж, защит, защищ, здоров, здрав, зрел, известн, изобил, изумл, изыскан, изю-*

минк, интерес, искрен, искрен, классич, комплимент, комфорт, конструктивн, красив, краснореч, красот, легк, ловк, лоялн, лучш, льгот, люб, любез, мил, мир, многофунк, молод, мораль, мудр, мужеств, наград, надеж, надежн, наслажд, нрав, обалден, обеспеч, обогат, обогащ, обожа, образов, обрат, обращ, одобрен, оживлен, опыт, остроум, ответствен, отзывчив, отлично, охотн, очаро, очищ, первоклас, плюс, побед, побежд, повыс, повыш, подвиг, поддерж, подлин, подробн, поздрав, поклон, полезн, положительн, польз, поощр, порядочн, потрясающ, похвал, почит, почтен, почтительн, прав, правд, праздн, превосход, преимуществ, прекрас, прелест, прибыль, приветлив, привлекат, прилич, приятн, проворн, продвинут, продуктив, проницательн, прост, процвет, прочн, пунктуальн, работоспособ, равномер, радова, радост, развле, разумн, рационал, реабилит, регулир, резви, рекоменд, респектабельн, реформ, самодостаточн, самоопредел, свеж, свобод, свят, слав, славн, смел, совмест, соглас, соответств, сотруднич, спас, сплочен, спок, способн, справедл, стабильн, старае, старал, старат, стараю, стойк, сторонник, стрем, счаст, творчес, точн, трогательн, убедит, уваж, уверен, увлека, удач, удив, удоб, удовлетвор, удовольст, украс, украш, улыб, умел, умн, ура, уравнове, усерд, усил, успех, успеш, утвер, утвердит, ухажив, харизм, хорош, храбр, цвести, цвет, целомудр, ценн, цивилиз, чемпион, честолюб, четк, чист, чуда, чуд, чудес, щедр, экономичн, элегантн, энергич, энтузиа, эрудир, этич, эффективн, ярк, ясн.

Словарь отрицательно окрашенной лексики включал в себя следующие элементы: *абсурд, аварии, авторитар, агрес, адск, аллергии, алчн, ампутир, анарх, аномал, антипатии, антироссийск, анти-социальн, апокалип, атак, банкрот, бедн, бедств, безмозг, безнадежн, безобраз, безраб, безум, беси, беспомощ, бессердеч, бесцельн, бесчувствен, беся, блокад, богохуль, бои, бой, бойкот, бойн, боле, болезн, боли, болн, болящ, бомб, боюсь, боя, варвар, взволнов, вздор, взлом, взятк, взяточничеств, вин, военн, возмут, возмущ, воин, войн, волнен, вопиющ, восстав, восстал, восстании, враг, враж, вред, вторг, вторж, выкин, высокомер, геноцид, глуп, гнев, гнусн, голод, горьк, грех, греш, груб, грущ, гряз, губит, двуличн, двусмыслен, демон, дерзк, дескримин, деспот, диссидент, душераздир, душил, душит, дьявол, ересь, еретик, еретич, ерунд, жадн, жадност, жалоб, жалов, жалуе, жертв, жестк, жесток, забит, заговор, запуг, затрудн, зверств, зло, злоупотребл, зомби, идиот, избив, избие, избил, избит, изгна, изгон, испуг, истощ, каприз, катастроф, категорич, кипен, клевет, клевещ, конфликт, коррупн, косо, кошмар, крадет, крадут, краж, краст, крут, лгал, лгать, лжи, ликвидир, лицемер, лишен, ловушк, ложн, ложн, ложь, мерзост, месть, меша, минус, мошен, мсти, мук, мучи, надое, напад, напад, нарко, наруш, наруш, насилии, нахал, не, неблаго, неверн, недостат, недоум, нелеп, ненави, ненормальн, необосн, непостоян, неправ, неприят, неприят, непродуктивн, нерв, несправедлив, несчаст, нетерп, неудоб, неэффектив, нищенск, нищет, обвин, обид, обиж, обман, обостр, ожог, опас, опас, оплак, осади, осажд, оскверн, оскорб, осужд, отврат, отвратительн, отвращ, отним, отнял, отнять, отомщ, отрав, отста, отсутств, отталк, отягчающ, ошиб, ошибк, падении, перебран, плачевн, плох, поджиг, поджог, подл, подожгл, позор, помех, поработи, порабоще, пораж, пориц, порок, пороч, порти, пострад, потряс, предав, предае, предат, предостер, прерыв, преслед, преступ, приниж, принуд, принужд,*

пристав, проблем, провал, провокац, провоцир, промах, пропаст, противн, противореч, прочь, пуга, пуглив, рабс, равнодуш, радикал, разби, раздраж, размыт, разорв, разочар, разруш, разрыв, расизм, расист, ошибс, падении, перебран, плачевн, плох, поджиг, поджог, подл, подожгл, позор, помех, поработи, порабоще, пораж, порищ, порок, пороч, порти, порча, пострад, потряс, предав, предае, предат, предостер, прерыв, преслед, преступ, приниж, принуд, принужд, пристав, проблем, провал, провокац, провоцир, промах, пропасть, противн, противореч, прочь, пуга, пуглив, рабс, равнодуш, радикал, разби, раздраж, размыт, разорв, разочар, разруш, разрыв, расизм, расист, удуш, ужас, укра, умер, умир, уничтож, уныл, унын, устаре, утрат, утрач, фальсифи, фанат, фашизм, фашист, хаос, хаотич, хваст, холод, худ, хуже, хулиган, черств, чужд, шантаж, шок, шум, эгоизм, эгоист, язв.

Все отрезки текста заносились в программу, позволяющую определить тональность текста. Анализ самой часто встречаемой в статьях категории «Украинский вопрос» показал следующие результаты.

Третьего февраля, когда категория впервые встречается в текстовом массиве, тональность текста равняется «–8». Далее эта же категория встречается четвертого февраля и ее тональность составляет «–20». 5-го февраля тональность текста данной категории падает до «–48». Седьмого февраля она достигает своего минимума за неделю «–83» и восьмого февраля снова поднимается до «–5». В текстах, опубликованных за время второй недели февраля, тональность составляла «–6» в первый день и далее снова опустилась до «–8». К концу недели она составляла «–13». Анализ третьей недели показал диапазон распределения сентимента от «–22» до «–74». В последнюю неделю февраля средняя тональность текстов, относящихся к категории «Украинский вопрос», составила «–44,2» и достигла своего минимума за все время наблюдений — «–93».

Стоит отметить, что анализируемый период всего лишь на два месяца отстоял от момента пикового противостояния на Украине (Майдан, декабрь 2014 г.), когда еще не до конца была понятна глубина происходящей трагедии и перспектива развития событий. Этим объясняется и использование в названии категории слова *вопрос*, семантика которого предполагает неопределенность и неясность.

Рис. 1. Результаты сентимент-анализа категории «Украинский вопрос»

Вторая по распространенности в тексте категория — «Отношения России с Китаем». Данная категория имеет ярко выраженную позитивную эмоциональную окраску. В первую неделю она находилась в диапазоне от 6 до 11. В текстах, опубликованных на третьей неделе, искомая категория отсутствует. В третью неделю она варьировалась от 38 до 27. И в четвертую неделю тональность составляла от 23 до 41. Таким образом, самая высокая тональность данной категории равна 41. Это показывает, что категория «Отношения России с Китаем» обладает самой высокой сентимент-оценкой. Здесь название категории также говорит само за себя — семантика слова *отношения* без атрибута предполагает *хорошие* отношения.

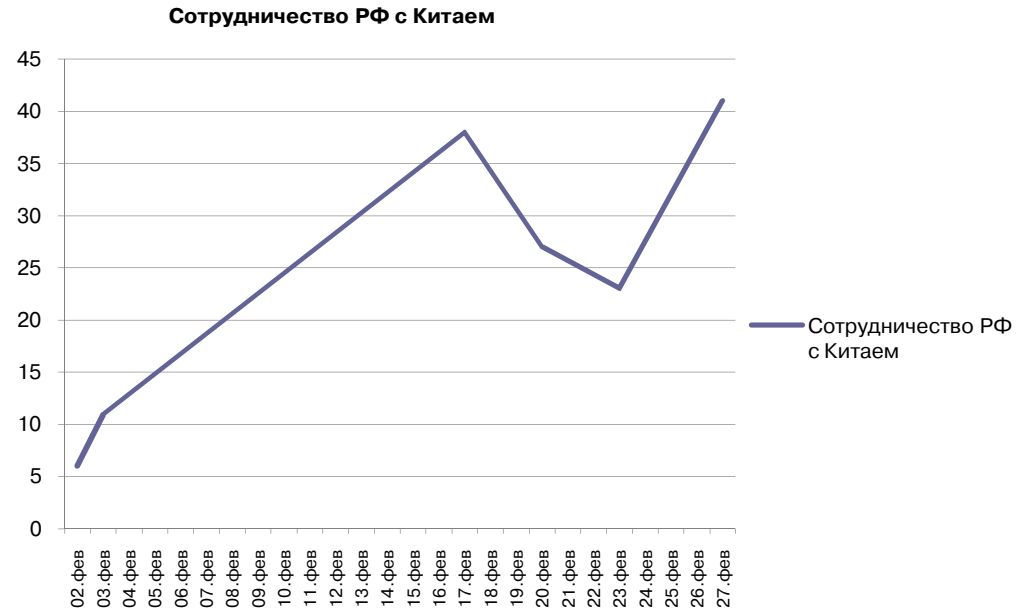

Рис. 2. Сентимент анализ категории «Сотрудничество РФ с Китаем»

Анализ тональности текста третьей по частоте упоминаний категории «Отношения России и Украины» показал, что тема эта имеет также исключительно негативную эмоциональную окраску. В первый день появления темы в тексте у нее была отмечена тональность текста «–3», во второй — «–6», в третий — «–5». Начиная со второй недели количественный показатель эмоциональной оценки текстов данной категории начинает заметно падать. Диапазон составляет от «–8» до «–18». На третьей неделе анализ показывает самый низкий уровень тональности текста — от «–10» до «–32».

Сентимент-анализ текстов, относящихся к категории «Конфликт в Сирии» показал, что все тексты в выборке имеют отрицательную тональность. На четвертой неделе исследований тональность данной категории достигает своего минимума за месяц. Он составляет «–23».

Рис. 3. Сентимент-анализ категории «Отношения РФ и Украины»

Рис. 4. Сентимент-анализ категории «Конфликт в Сирии»

Далее идет категория «Сотрудничество с Туркменистаном». Тональность текстов данной категории является только позитивной. Наивысшую оценку тональности текста здесь получил текст, опубликованный на сайте в последнюю неделю исследования, она составляет «27».

Тексты категории «Отношения России и Греции» также окрашены позитивно. Самая высокая оценка равна «9».

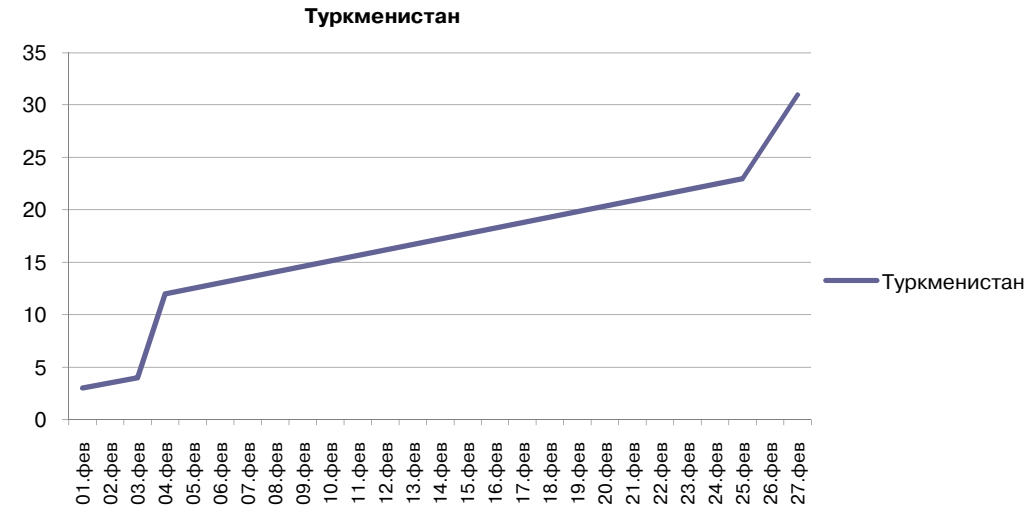

Рис. 5. Сентимент-анализ категории «Туркменистан»

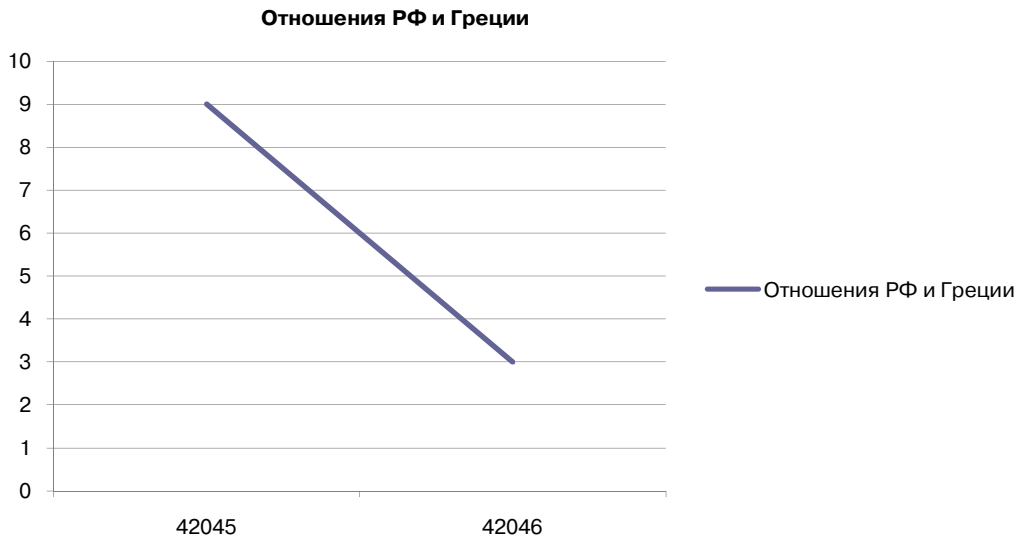

Рис. 6. Сентимент-анализ категории «Отношения РФ и Греции»

Категория «Санкции против России» предсказуемо имеет негативную эмоциональную окраску. В первую неделю сентимент равен «–4», начиная с третьей недели оценка начинает резко падать. В текстах третьей недели февраля она составляет от «–35» до «–53». На четвертой неделе эмоциональная окраска текстов данной категории снова понижается и располагается в диапазоне от «–9» до «–41».

Рис. 7. Сентимент-анализ категории «Санкции против России»

Последней проанализированной категорией была «Дипломатия сегодня». Впервые данная тема поднимается в текстах второй недели. Эмоциональная окраска составляет от «4» до «9». В текстах, опубликованных на сайте на третьей неделе февраля, тема дипломатии не поднималась.

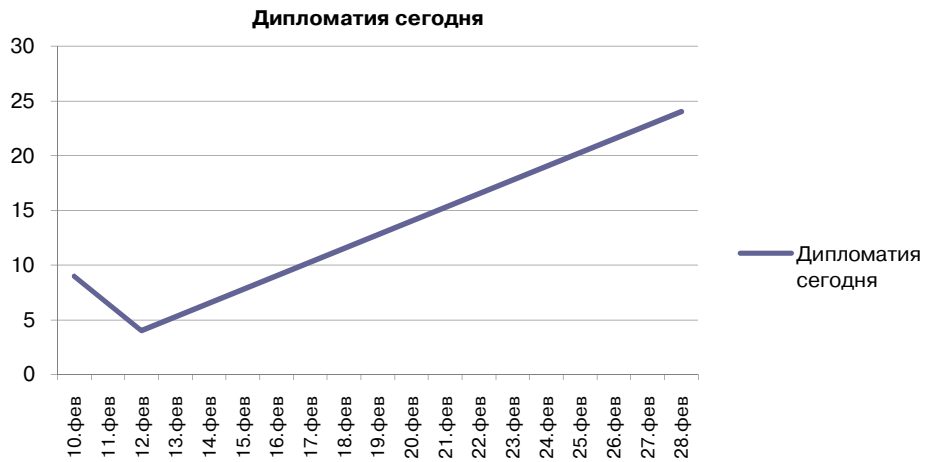

Рис. 8. Сентимент-анализ категории «Дипломатия сегодня». Результаты проведенного сентимент-анализа

Проведенный сентимент-анализ текстов сайта МИД РФ показал, что самые популярные категории чаще имеют положительную эмоциональную окраску. Самую низкую тональность имеют тексты, относящиеся к тематической категории «Украинский вопрос». Кривая, показывающая тональность текста данной категории на графике, опускается до отметки «–90» и не поднимается выше «6». Вто-

рая по объему занимаемого текста в статьях категория «Сотрудничество России с Китаем», напротив, имеет позитивную окраску. Сентимент-анализ показал, что самая высокая тональность текста в один из дней составляла «41», самая низкая — «6». В категории «Отношения России с Украиной» присутствует значительное число негативно окрашенной лексики. Тональность находится в диапазоне от «–3» до «–25». Напротив, тексты категории «Сотрудничество России с Туркменистаном» отличаются сравнительно большим количеством позитивно окрашенной лексики. Оценка здесь варьируется от «3» до «31». Тексты категории «Отношения России и Греции» в феврале также были окрашены положительно — от «+3» до «+9». Категория «Санкции против России» выделяются низкой оценкой тональности — от «–4» до «–53». Показатели категории «Дипломатия сегодня» по своей тональности варьируются от «4» до «24».

Таким образом, положительно окрашенными оказались категории «Сотрудничество России с Китаем», «Сотрудничество России с Туркменистаном», «Отношения России и Греции» и «Дипломатия сегодня». Негативную окраску имеют следующие категории: «Украинский вопрос», «Отношения России с Украиной» и «Санкции против России».

ЗАКЛЮЧЕНИЕ

Один и тот же корпус текстов сайта МИД России, размещенных в разделе «Новости», был проанализирован двумя наиболее популярным сегодня формальными методами — контент- и сентимент-анализом. Контент-анализ [Беляков 2016] показал, что большее внимание уделяется теме «Украинский вопрос», второй по объему занимаемого текстового пространства является тема «Сотрудничество России с Китаем», далее идет тема «Конфликт в Сирии». После нее идет тема «Сотрудничество с Туркменистаном», далее — «Отношения России и Греции», после нее «Санкции против России» и далее тема «Дипломатия сегодня». Сентимент-анализ показал, что из них положительно окрашенными оказались категории «Сотрудничество России с Китаем», «Сотрудничество России с Туркменистаном», «Отношения России и Греции» и «Дипломатия сегодня». Негативную окраску имеют следующие категории: «Украинский вопрос», «Отношения России с Украиной» и «Санкции против России».

Сентимент-анализ — один из современных методов анализа текстов, используемый в прикладной лингвистике. За последние годы даже открылось немало коммерческих фирм, предлагающих свои услуги по анализу тональности текстов, что необходимо в первую очередь для анализа отзывов покупателей о приобретенных товарах. Все это перспективные направления в лингвистических исследованиях, используемые как производителями, заинтересованными как можно выгоднее продать свой товар, так и политическими и государственными структурами.

Анализ тональности за последние годы стал одним из самых перспективных направлений компьютерной лингвистики, в результате чего за последние 10 лет появилось немало работ в этом направлении, однако лишь немногие из них концентрировались на методах глубинного анализа тональности, который, несо-

мненно, требует дальнейших исследований. Применение систем анализа тональности может в значительной степени упростить процесс анализа мнений о событиях, происходящих в мире, так как при помощи анализа тональности становится возможным создание систем автоматической оценки общественного мнения [Максименко 2014].

Дата поступлекния: 24.04.2016
Дата принятия к печати: 07.06.2016

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Беляков М.В.* (2016). Анализ новостных сообщений сайта МИД РФ методом контент-анализа (статья 1) [*Belyakov M.V.* Content-analysis of the site mid. ru news] // Вестник РУДН. Серия: Теория языка. Семиотика. Семантика. № 3. 2016. С. 58—67.
2. *Максименко О.И.* (2014). Анализ тональности текстов (сентимент-анализ) на материале текстов СМИ [Maksimenko O.I. Sentiment-analysis of media texts] // IV Новиковские чтения: Функциональная семантика и семиотика знаковых систем. Сб. научных статей. Часть 1. М.: РУДН. С. 96—105.

THE ANALYSIS OF NEWS MESSAGES OF THE RF MINISTRY OF FOREIGN AFFAIRS WEB-SITE BY THE SENTIMENT-ANALYSIS

(article 2)

M.V. Belyakov

Moscow State Institute of Foreign Relations (University)
76, Vernadsky ave., Moscow, Russia, 119454
belmax007@hotmail.com

Text sentiment-analysis represents one of the most actual and modern problems of the applied linguistics. The sentimant-analysis, being one of formal methods of the text analysis the, allows to estimate, on the one hand, emotional tone of the text, and on the other to draw substantial conclusions by results of revealed sentiment. In the era of the information wars this task is the task of vital importance. The article considers the variant of the sentiment-analysis by the example of some categories generated on materials of texts of a site the Ministry of Foreign Affairs of the Russian Federation for one month. To carry out of the analysis 2 dictionaries — positively and negatively marked words have been received. Results of processing by the program of the sentiment-analysis have allowed to allocate the text and graphically present an estimation of texts tonality of the categories that has confirmed an opportunity of using this formalized method for public opinion estimation and formation.

Key words: diplomatic discourse, web-site, sentiment-analysis, MID RF