УДК 004.912

АНАЛИЗ ЭМОТИВНОСТИ ТЕКСТОВ НА ОСНОВЕ ПСИХОЛИНГВИСТИЧЕСКИХ МАРКЕРОВ С ОПРЕДЕЛЕНИЕМ МОРФОЛОГИЧЕСКИХ СВОЙСТВ

Д. В. Гудовских, И. А. Молошников, Р. Б. Рыбка

Национальный исследовательский центр «Курчатовский институт»

Поступила в редакцию 14 апреля 2015 г.

Аннотация: *в статье предложен подход, целью которого является определение эмоционального состояния автора текста на русском языке с использованием психолингвистических маркеров эмотивности, оцениваемых на основе морфологических характеристик слов. Представленный подход позволяет выделить эмотивные тексты в тематической коллекции, отражающие возбужденное состояние автора в момент написания и, как следствие, может быть основой для систем оценки социальной напряженности на основе интернет-источников.*
Ключевые слова: *анализ текстов, психолингвистические маркеры, эмоционально окрашенный текст.*

Abstract: *the paper addresses the issue of determination of the emotional state of the author. The research is based on the Russian texts. The methodemploys morphological properties of words to define psycholinguistic markers of emotivity. The approach allows for allocating emotive texts that reflect their authors' emotional arousal. One of the practical applications of the method is evaluation of social tension in computer-mediated communication.*
Key words: *text mining, psycholinguistic markers, emotive text.*

С ростом популярности и доступности социальных сетей в мире одной из актуальных задач анализа больших данных является анализ неструктурированного текста с целью извлечения и оценки выражаемого мнения автора. Этой тематике посвящено значительное число работ, в которых авторы предлагают различные подходы к решению задач сентимент анализа текста и оценки мнения автора [1]. Эти системы продемонстрировали свою применимость в различных областях деятельности человека, таких как реклама, маркетинг, социология, политология, криминалистика и т.д. Для каждой области есть свой специфический уклон. Так, эксперты психолингвистики в первую очередь решают проблемы психотипического плана [2–3], судебная экспертиза занята решением вопросов установления авторства и условий написания текста [4] и т.д. С помощью автоматизированных систем оценивается тональность текстов, идентифицируются такие признаки, как пол автора текста, особенности его психотипа, его возраст, профессия, родной язык, место формирования языковых навыков и т.д. [5], но они не оценивают психологического (эмоционального) состояния автора текста в момент его написания. В связи с этим в данной работе была поставлена цель построения методики оценки эмоционального состояния автора текста с использованием психолингвистических маркеров эмотивности, основывающихся на морфологических признаках.

Дополнение существующих автоматизированных систем анализа данных социальных сетей средствами оценки эмоционального состояния автора текста позволило бы расширить их возможности и решать задачи по предупреждению актов суицида, оценки эмоционального состояния общества, выявлению периодов социально-экономических волнений (кризисов).

Задача определения маркеров оценки состояния автора текста активно изучается специалистами в области психолингвистики. Так, в работе [3] исследователи из России показали корреляцию количественных признаков текстов и эмоционального состояния автора. При этом использовались психологические тесты для оценки состояния испытуемых, далее по текстам, написанным в данном состоянии, подсчитывали различные синтаксические и морфологические признаки. Среди основателей психолингвистики, подробно изучавших вопросы языка с учетом психологических особенностей автора, стоит отметить труды таких международных экспертов, как Ч. Осгуд, Н. Хомский, Дж. Миллер, а также представителей Русской школы Л. С. Выготского и А.Н. Леонтьева [6].

Описание подхода

Для анализа текстов из интернет-источников, имеющих свою специфику, первоначально были вы-

браны некоторые психолингвистические маркеры из числа способных отражать степень эмоционального напряжения. Таким образом, выделяются стилистические особенности текста, характеризующие эмоциональное состояние автора:

– количество слов в тексте;

– средний размер предложений в словах;

– соотношение количества глаголов к количеству существительных в единице текста;

– количество знаков восклицания в документе;

– наличие эмотиконов (символов, символизирующих эмоции).

На основании результатов предыдущего исследования применения психолингвистических маркеров, было установлено, что некоторые коррелирующие маркеры позволяют отделить среди набора текстов объективные тексты с фактами и новостями от субъективных текстов, выражающих мнение автора. Наиболее сильно отражают эмоциональную возбужденность и соответственно используются в данной работе следующие маркеры:

1. Коэффициент Трейгера (КТ) – соотношение количества глаголов к количеству прилагательных в единице текста. Нормальное значение близко к 1.

2. Коэффициент определенности действия (КОД) – соотношение количества глаголов к количеству существительных в единице текста. Нормальное значение также близко к 1.

3. Коэффициент агрессивности (КА) – отношение количества глаголов и глагольных форм (причастий и деепричастий) к общему количеству всех слов. Нормальное значение не превышает 0,6.

Выбранные маркеры имеют общие характеристики при диагностировании:

– завышенные значения указывают на наличие эмоционального беспокойства, что в целом характерно для лиц, склонных к активным действиям;

– низкие значения указывают на такие личные характеристики, как неуверенность, зависимость, тревога.

В работе рассматривались некоторые основные источники данных четырех типов.

1. Новостные источники – включают в себя тексты от СМИ.

2. Социальные сети – платформы, сервисы или сайты, предназначенные для построения, отражения и организации социальных взаимоотношений в сети. Содержат тексты различного характера: новостные ленты, отзывы, комментарии пользователей.

3. Блоги – это постоянно поддерживаемые сайты. Обычно блоги являются тематическими, отражающими точку зрения автора. Основное содержание блога постоянно обновляется и состоит из текстов различной длины.

4. Микроблоги – то же самое, что и блог, но имеет ограничение размера сообщения (100–200 символов).

Процесс оценки эмотивности документа делится на следующие этапы:

1) предобработка: на данном этапе текст разбивается на предложения и очищается от несмысловых частей;

2) выделение признаков: текст проходит обработку морфологическим модулем и блоком снятия морфологической омонимии (определяются части речи для слов);

3) оценка документа: рассчитываются значения маркеров.

Важным этапом оценки эмотивности текста является определение частей речи членов предложения. Среди существующих открытых решений для морфологического разбора русскоязычного текста можно выделить АОТ, Mystem и Freeling. В данной работе использовался Mystem [7]. Его работа основана на использовании словаря Зализняка. На выходе библиотека выдает все возможные варианты морфологического разбора данного слова с вероятностями его использования в данном контексте. Точность работы разборщика Mystem 3.0 мы оценивали на текстах с однозначной морфологической разметкой из СинТагРус, части Национального корпуса русского языка (НКРЯ) [8]. Оценка проводилась с использованием встроенного в Mystem алгоритма снятия омонимии. Существующие различия в формате представления морфологических признаков программой Mystem и форматом НКРЯ устранялись функцией преобразования результатов разбора Mystem к формату НКРЯ. Разбор считался правильным при установлении взаимно однозначного соответствия. Оценка работы встроенного в Mystem механизма снятия морфологической неоднозначности (омонимии) демонстрирует точность около 50 %, что требует его улучшения для качественного анализа. Оценка полноты всех возможных вариантов разбора слов в Mystem в соответствии со значениями НКРЯ показала, что возможных правильных вариантов разбора существует для 94 % используемых в корпусе слов. Для улучшения работы морфологического разборщика был разработан механизм снятия морфологической омонимии на основе решения классификационной задачи с помощью машины опорных векторов (SVM). Следует отметить, что развитие методики оценки потребует расширения набора маркеров более высокого уровня, для извлечения которых необходим полный морфологический разбор слов, а не только определение части речи.

В рамках этой задачи предложение рассматривается как последовательность слов со словоформами $\{w1.wN\}$, в котором для всех слов известны все возможные варианты полного морфологического разбо-

ра – теги. Знаки препинания также учитываются как отдельные слова и помечаются единственным общим тегом PUNC. Последовательный алгоритм обработки предполагает обработку предложения справа налево, т.е. с конца, в котором на каждом *i*-м шаге учитываются все известные признаки в том числе уже разобранных слов. Для каждого слова w_i формируется вектор признаков, в который входят его значения и значения ближайших соседей из некоторого окна *W*. В экспериментах использовалось подобранное опытным путем окно из восьми слов *W* = (*i*-3, *i*-2, *i*-1, *i*, *i*+1, *i*+2, *i*+3, *i*+4).

Вектор признаков включает следующую информацию для каждого слова:

1) словоформы всех слов из окна *W*;

2) теги для тех слов из *W*, для которых они уже проставлены;

3) классы неоднозначности для всех слов из *W* (+ их биграммы и триграммы), т.е. совокупность всех возможных тегов для слова. Мы представляем его в виде конкатенации строк тегов. Например, для слова «Эти» из предложения-примера класс неоднозначности выглядит так: А|ИМ|МН_А|ВИН|МН|НЕОД;

4) возможные теги для всех слов из W;

5) подтеги, т.е. отдельные морфологические признаки, для тех слов из *W*, для которых теги уже проставлены;

6) возможные подтеги для всех слов окна.

Были выделены 43 морфологических признака в соответствии со значениями НКРЯ. На каждый признак обучалась бинарная классификационная модель. При бинарной классификации реализации SVM в качестве результата работы выдают не только выбранный класс, но и действительное число – decision value (dv), которое позволяет оценить, насколько близок данный экземпляр к классу. Пусть для данного слова *I* – пересечение множеств подтегов его возможных тегов, а *U* – объединение множеств подтегов его возможных тегов, *C* – множество подтегов его правильного тега. Таким образом, мы преобразуем данные окна *W* в точку *x*, и для классификаторов, соответствующих подтегам, входящим в $(U\setminus I)\cap C$, отнесем *x* в класс «правильно», а для входящих в $(U\setminus I)\setminus C$ – в класс «неправильно». При разметке выбираем из возможных тегов тот (обозначим его множество подтегов также через *C*), для которого максимальна сумма decisionvalue по всем классификаторам, соответствующим возможным подтегам:

$$\sum_{(I\setminus U)\cap C} Decision\ value + \sum_{(I\setminus U)\setminus C} Decision\ value \to \max$$

Экспериментальная часть

Тестирование работы морфологического теггера

Тестирование модуля морфологического разбора проводилось на корпусе текстов из НКРЯ. Обрабатываемый корпус содержит 366 245 слов. Для обучения расчетных моделей были использованы 90 % случайно выбранных слов, а на оставшиеся 10 % (38 959 слов) проводилось тестирование. Ниже в таблице представлены результаты тестирования определения частей речи.

Таблица

Результаты тестирования

Название признака	Количество в выборке	Тестовое количество	Точность определения данного класа (recall)
ADV	23 625	2363	0,96
CONJ	23 450	2366	0,95
NUM	3118	299	0,98
PART	17 556	1707	0,92
PR	39 007	3890	1,00
S	158 257	15 771	0,99
V	61 879	6343	0,99
A	61 372	6107	0,96
INTJ	59	4	0,50
COM	207	23	0,17

Очевидно, что некоторые классы (COM, INTJ) слабо представлены в выборке, следовательно, и результаты по классификации их низкие. Из-за нехватки примеров в выборке было принято решение обучить модели на полном наборе данных корпуса. Тестирование собранного программного модуля полного морфологического разбора слов с использованием 43 классификаторов и оценкой наиболее вероятного из вариантов, предоставленных Mystem, показало точность определения 93,93 %. Достигнутая точность совпадает с полученным ранее значением всех возможных вариантов Mystem для корпусных слов.

Эксперименты с маркерами эмотивности

Для исследований была собрана выборка текстов на тему запуска ракет «Булава» (система Brand analytics [9]).

Исследуемая выборка состоит из 8503 сообщений на тему запуска ракет «Булава», собранных в период с 8 сентября по 17 октября 2014 г. Она разделена по типам источников с целью выделить особенности дневных показаний маркеров. Из событий, зафиксированных за данный период, знаковым служит удачный запуск ракеты 10 сентября. Предварительный анализ маркеров КОД, КТ, КА показал, что более показательными являются значения стандартного отклонения маркеров с точки зрения выделения повышенного эмоционального фона дня. Стандартное отклонение значений маркеров отражает уровень изменчивости эмоционального напряжения (фона) в высказываниях на заданную тему. На рис. 1, 2, 3 отражены показатели среднего отклонения для КОД, КТ и КА соответственно.

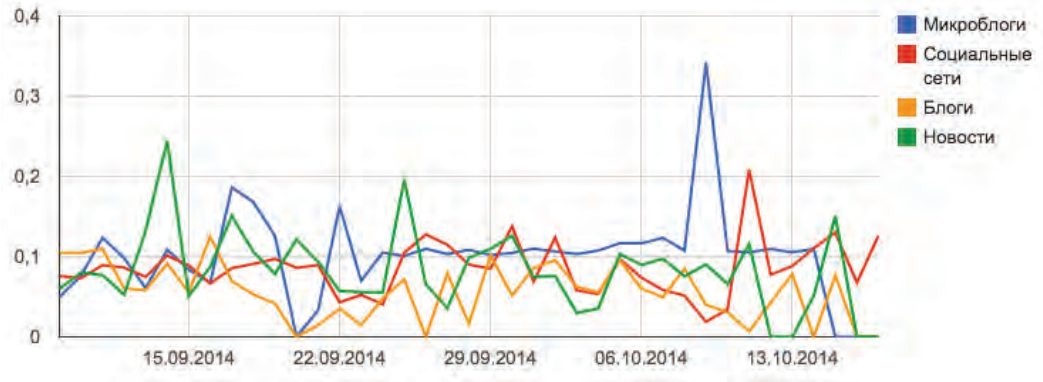

Рис. 1. Среднее отклонение КОД

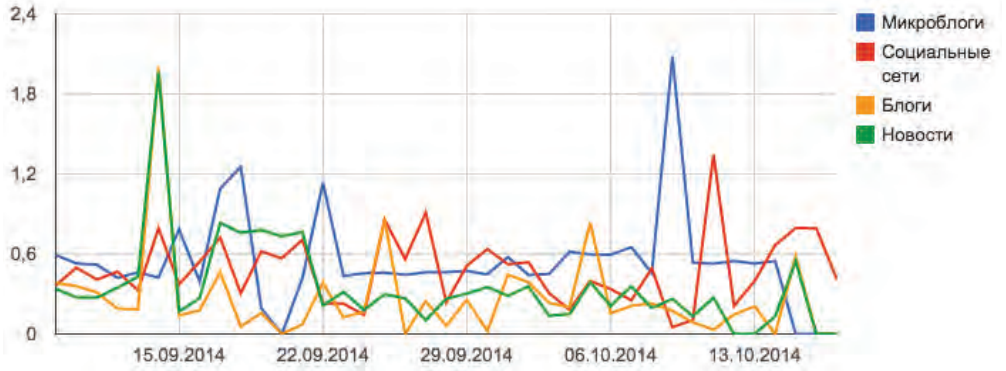

Рис. 2. Среднее отклонение КТ

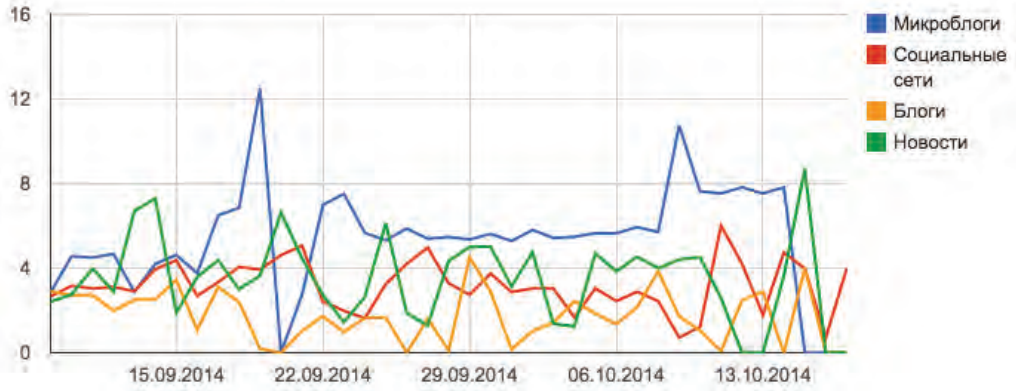

Рис. 3. Среднее отклонение КА

Наблюдается общая зависимость маркеров во всплесках отклонения на всем периоде. Стоит отметить сильные колебания для новостных ресурсов, что не проявлялось на значениях средних показателей. На основе данного графика выделяются следующие стандартные отклонения: для микроблогов – 0,1, социальных сетей и новостей – 0,08, для блогов – 0,05. Интересные активности проверялись вручную экспертом:

– 14 сентября – всплеск связан с эмоциональной активностью в обсуждении результатов запуска прошедшей недели и возникших слухов о повторных запусках;

– рост показателей 16–17 сентября связан с регулярно обсуждаемой в новостях темой второго пуска ракеты «Булава» (после заявления заместителя министра обороны России о точной дате второго запуска);

– 25 сентября – упоминание «Булавы» появилось в украинских источниках новостей с эмоционально выраженным антироссийским контекстом;

– рост показателей в социальных сетях с 24 по 26 сентября связан с широко обсуждаемой темой ядерного противостояния США и России. На пике 26 наблюдается увеличение неновостных сообщений, выражающих мнение авторов.

В целом показания и особенности КТ совпадают с КОД, но стоит отметить более ярко выраженные отклонения данного коэффициента для блогов. Выделяются следующие стандартные значения отклонений: для микроблогов – 0,5, социальных сетей – 0,45, новостей – 0,2, блогов – 0,3. В целом, пики совпадают с выделенными КОД, но есть и отличия в выделенных днях для сообщений в блогах:

– за 14 и 25 сентября для блогов выделены более сильные отклонения по сравнению с показаниями КОД. В данные дни наблюдались достаточно агрессивные сообщения пользователей;

– в своих минимумах 20, 26, 30 сентября среди постов в социальных сетях не встречается эмоциональных высказываний и обсуждений; в основном в эти дни новостные тексты.

Пик за 12 октября в социальных сетях связан с активными дебатами на тему военного противостояния США и России.

По графику выделяются следующие стандартные уровни отклонения коэффициента агрессивности: для микроблогов – 5, социальных сетей – 3, новостей – 3,4, блогов – 1,5.

Всплески 13, 14, 21, 25 сентября, а также с 5 по 9 октября связаны, не выделяют эмотивных текстов, из чего следует, что стандартное отклонение до 10 не является явным указанием на повышенный эмоциональный фон дня.

21, 27 сентября, 11 октября в социальных сетях обнаружены восклицательные-эмотивные отзывы, из чего мы сделали вывод о том, что стандартное значение отклонения для социальных сетей равно 4 и наиболее показательным для комментариев в социальных сетях является коэффициент агрессивности.

Исходя из анализа дневных значений среднего отклонения маркеров выделим следующие особенности:

1) среднее отклонение является более показательным, чем средние дневные значения;

2) близость значений маркеров указывает на повторяющийся контент сообщений;

3) заниженные пороговые значения маркеров по сравнению с указанными в психолингвистической литературе;

4) отсутствие взаимосвязи отклонений дневных значений маркеров для сообщений в микроблогах и эмотивности текстов.

Особенности 3 и 4 связаны с объемами анализируемых сообщений, поскольку при психолингвистическом анализе требуемый объем текста равен 150–200 словам. Сопоставимый объем в большей степени обеспечивается авторами на новостных ресурсах, блогах и социальных сетях. При анализе микроблогов соответствия данному требованию можно добиться анализом ряда сообщений отдельного автора за период по данной теме, объединенных в один текст.

Результаты тестирования представленной системы оценки эмотивности текстов показывают, что выбранные психолингвистические маркеры выделяют сильную эмоциональную реакцию автора, выраженную в тексте, а дневные показания эмотивности отражают уровень социальной напряженности в сети касательно данной темы.

Таким образом, в работе представлена система оценки эмотивности текстов на русском языке, которая позволяет исследовать уровень эмоциональности социальной реакции относительно конкретного события. Исследование проводилось на ограниченной выборке текстов из 4 типов интернет-источников, собранных за определенный период. Оно продемонстрировало зависимость дневных значений стандартного отклонения выделенного набора психолингвистических маркеров и эмотивно окрашенных текстов. Результаты показывают, что выбранные психолингвистические маркеры выделяют сильную эмоциональную реакцию автора, выраженную в тексте, а дневные показания эмотивности отражают уровень социальной напряженности в сети касательно данной темы. Была исследована взаимосвязь эмотивности текста и конкретного события (запуск ракеты «Булава»). Как упоминалось ранее, предлагаемый подход может служить в качестве дополнительного аналитического инструмента, на основе которого можно строить оценку социально-эмоционального фона тематики в Интернете. Выявлена необходимость адаптации разработанной методологии отдельно для таких источников, как микроблоги, благодаря особенности коротких сообщений. В связи с важностью задачи оценки сообщений в микроблогах адаптация представленной методики для анализа коротких сообщений пользователей – одна из целей нашего дальнейшего развития системы. Второй целью дальнейших работ является расширение набора маркеров с использованием полного морфологического разбора слов.

ЛИТЕРАТУРА

1. *Pang B.* Opinion mining and sentiment analysis / B. Pang, L. Lee // Foundations and Trendsin Information Retrieval. – 2008. – Vol. 2. – № 1/2.

2. *Чернов Д. Н.* Выражение психологических особенностей в количественных показателях речи / Д. Н. Чернов, Ю. Ю. Игнатов // Вопросы психолингвистики. – 2013. – № 1 (15).

3. *Литвинова Т. А.* Частоты встречаемости последовательностей частей речи в тексте и психофизиологические характеристики его автора : корпусное исследование / Т. А. Литвинова, О. А. Литвинова, П. В. Середин // Вестник Иркут. гос. лингв. ун-та. – 2014. – № 2. – С. 8–12.

4. *Olsson J.* Forensic Linguistics, Second Edition / J. Olsson. – London : Continuum, 2008.

5. *Peersman C.* Predicting Age and Gender in Online Social Networks. Proceedings of the 3rd international workshop on Search and mining user-generated contents. – New York, USA, 2011.

6. *Леонтьев А. А.* Основы психолингвистики / А. А. Леонтьев. – М. : СМЫСЛ, 1997. – 287 с.

7. *Gudovskikh D.* Automatic selection psycholinguistic characteristics of texts in the concept of Big Data / D. Gudovskikh. – 2013.

8. Mystem. – Mode of access: https://tech.yandex.ru/mystem/

9. Russian National Corpus. – Mode of access: http://www.ruscorpora.ru/en/index.html

10. Система Brand Analytics, компания «Ай-Теко». – Режим доступа: http://br-analytics.ru/

Национальный исследовательский центр «Курчатовский институт»

Гудовских Д. В., инженер-исследователь, аспирант

E-mail: dmitrygagus@gmail.com

Тел.: 8-963-994-26-50

Молошников И. А., инженер-исследователь, аспирант

E-mail: ivan-rus@yandex.ru

Тел.: 8-926-378-09-57

Рыбка Р. Б., инженер-исследователь

E-mail: rybkarb@gmail.com

Тел.:8-926-344-61-35

National Research Center «Kurchatov Institute»

Gudovskikh D. V., Research Engineer, Post-graduate Student

E-mail: dmitrygagus@gmail.com

Tel.: 8-963-994-26-50

Moloshnikov I. A., Research Engineer, Post-graduate Student

E-mail: ivan-rus@yandex.ru

Tel.: 8-926-378-09-57

Rybka R. B., Research Engineer

E-mail: rybkarb@gmail.com

Tel.:8-926-344-61-35