ВЕСТНИК НАЦИОНАЛЬНОГО ИССЛЕДОВАТЕЛЬСКОГО ЯДЕРНОГО УНИВЕРСИТЕТА "МИФИ", 2020, том 9, № 3, с. 279–288

ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА

УДК 004.91

КОМПЛЕКСНЫЙ ИНСТРУМЕНТ ДЛЯ АВТОМАТИЗИРОВАННОГО ТОНАЛЬНО-ЭМОТИВНОГО АНАЛИЗА ТЕМАТИЧЕСКИХ ТЕКСТОВ

© 2020 г. А. В. Наумов[1,*], А. А. Селиванов[1,], И. А. Молошников[1,***], А. Г. Сбоев[1,2,****]**

[1] *Национальный исследовательский центр "Курчатовский институт", Москва, 123182, Россия*

[2] *Национальный исследовательский ядерный университет "МИФИ", Москва, 115409, Россия*

**e-mail: sanya.naumov@gmail.com*

***e-mail: aaselivanov.10.03@gmail.com*

****e-mail: dmitrygagus@gmail.com*

*****e-mail: sag111@mail.ru*

Поступила в редакцию 23.06.2020 г.
После доработки 30.06.2020 г.
Принята к публикации 08.09.2020 г.

В работе представлен инструмент для комплексного тематического, аспектно-тонального и эмотивного анализа текстовых данных написанных на естественном языке. Для тематического анализа применяется метод на базе вероятностно-энтропийных характеристик, позволяющий автоматически определять количество тематик в текстовой коллекции. Аспектно-тональный анализ реализуется на базе нейросетевой модели сложной топологии – интерактивной сети внимания. Апробация модели на корпусе соревнования SentiRuEval-2015 продемонстрировала точность 0.58 по метрике f1-macro, что выше по сравнению с существующими результатами. Для проведения эмотивного анализа создан метод на базе контекстно-зависимых векторных представлений слов с последующей обработкой ансамблевым классификатором. Обучение проводилось на специально подготовленном корпусе предложений для 5 базовых эмоций (радость, грусть, злость, страх и удивление). Предложения для корпуса собраны из различных типов текстов и размечены средствами краудсорсинга. Точность метода составила 0.76 по метрике f1-macro. В статье также продемонстрирован пример использования выбранного комплекса инструментов. Для анализа выбраны тексты социальной сети LiveJournal и новостей из проекта SCTM-ru. Представлена визуализация результатов анализа текстов в виде графов, которая показывает эффективность созданного комплекса инструментов.

Ключевые слова: анализ данных, текстовые данные, обработка естественного языка, аспектный сентимент анализ, выделение эмоций, нейронные сети, машинное обучение

DOI: 10.1134/S2304487X20030086

1. ВВЕДЕНИЕ

Рост количества открытой информации в интернет-среде вызывает потребность развития автоматизированных инструментов ее разноуровневого анализа. Одним из ключевых видов информации при этом являются тексты на естественном языке, в анализе которых заинтересованы специалисты и аналитики различных предметных областей, в том числе: социологии, психологии, криминалистики, политологии, лингвистики, конфликтологии, которым необходимы решения для анализа потоков документов, способные извлекать признаки тонально-эмотивного отношения людей к различным тематикам и объектам.

На текущий момент известны несколько систем комплексного анализа данных, в том числе: IBM Watson Explorer, iFORA, Семантический архив [1–3]. Они представляют решения для информационного поиска, определения именованных сущностей, инструментов работы с тематиками и фактами. Однако комплексных систем текстового анализа, позволяющих оценить не только тональную составляющую текста, но и выраженную в нем эмотивную реакцию автора, в литературе не представлено.

Наиболее перспективными методами для решения задач аспектно-тонального анализа (когда тональность определения для пары: аспект–контекст) являются нейросетевые методы с различными механизмами внимания [4, 5] и векторными представлениями слов [6, 7]. Аспектом в данном случае может быть заранее заданный объект в виде фразы из текста или определенной категории. Что касается определения эмоций автора в тексте, то для корпусов на иностранных языках, наилучшие результаты показывают ансамбли методов машинного обучения в совокупности с век-

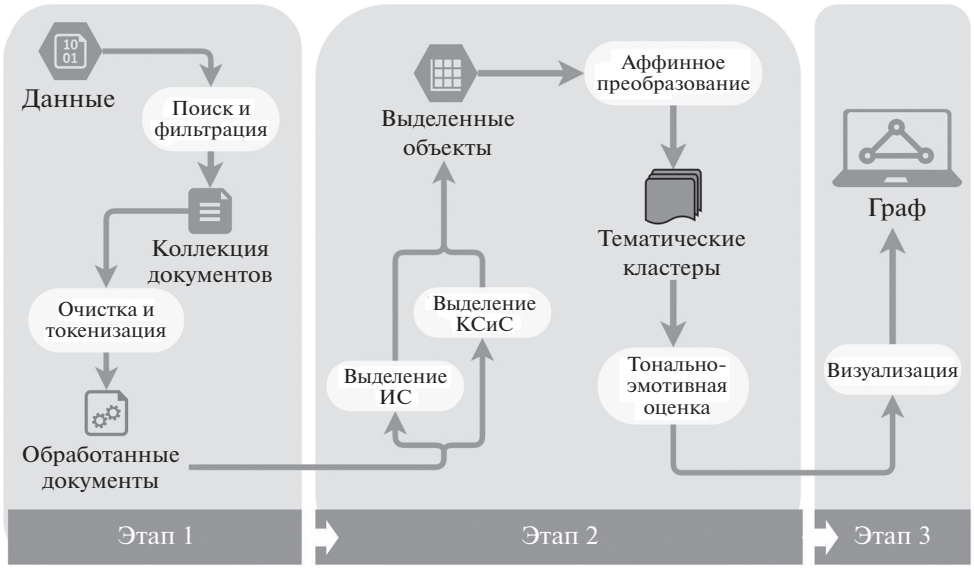

Рис. 1. Общая схема комплексного инструмента тонально-эмотивного анализа. На рисунке использованы следующие аббревиатуры: ИС – именованная сущность, КСиС – ключевые слова и словосочетания.

торными представлениями слов [8, 9] и словарями эмотивной лексики [10, 11]. Для русского языка отсутствуют достаточные наборы размеченных примеров для задачи определения эмоций, что является дополнительной трудностью.

В связи с этим целью данной работы является разработка комплексного инструмента для автоматизированного тонально-эмотивного анализа текстов заданной тематической направленности. Предлагаемый инструмент включает в себя алгоритмы: а) вероятностно-энтропийного анализа текстов для выделения тематик, б) аспектного анализа тональности, на основе глубокой нейронной сети с вниманием, где в качестве аспектов рассматриваются именованные сущности категорий Персона\Локация\Организация, в) эмотивного анализа с выделением эмоций автора, выраженных в тексте, с помощью моделей машинного обучения, а также г) средств визуализации результатов в виде аспектно-тональных семантических графов.

2. ВЫБРАННЫЕ ИНСТРУМЕНТЫ АНАЛИЗА ТЕКСТОВЫХ ДАННЫХ

2.1. Этапы обработки текстовых данных

Описание предложенной методики автоматизированного тонально-эмотивного анализа текстов заданной тематической направленности, написанных на естественном языке, представлено на рис. 1.

2.2. Предобработка текстовых данных

Исходный набор данных преобразуется в необходимый формат путем нескольких процедур: выбор коллекции документов, очистка текста от тегов разметки и других бессмысленных выражений, разбиение текста на отдельные слова и предложения, а также определение леммы для каждого слова.

Коллекция документов формируется в зависимости от задачи на основе текстов, выбранных для анализа. Например, это могут быть документы определенной тематики, или документы содержащие определенные слова.

Очистка текста происходит с помощью регулярных выражений. При этом удаляются html-тэги и адреса сайтов.

Очищенный текст разбивается на слова и предложения с помощью библиотеки с открытым исходным кодом UDpipe[1] для обработки текстов на естественном языке, которая является универсальной, с поддержкой многих языков (в т.ч. русского) и показывающей высокие результаты в сравнении с другими подобными средствами [12]. Точность лемматизации, выделения слов и предложений для русского языка составляет 96.6%, 99.7% и 98.8%[2] соответственно на корпусе СинТагРус.

[1] Сайт библиотеки – http://ufal.mff.cuni.cz/udpipe
[2] Модель 2.5: http://ufal.mff.cuni.cz/udpipe/models#universal_dependencies_25_models_description

2.3. Тематический анализ

Подготовленные на предыдущем этапе документы анализируются на предмет выделения тематических кластеров и извлечения именованных сущностей.

Для выделения тематических кластеров был выбран вероятностно-энтропийный подход, описанный в работе [13], основанный на широком наборе энтропийных характеристик текста, сочетающем в себе различные распределения слов в коллекции, как TF-IDF, дивергенция Кульбака–Лейблера, информационной энтропии, семантического алгоритма Гинзбурга и т.д., с последующей кластеризацией на базе аффинного преобразования. При этом тематика характеризуется набором ключевых слов и словосочетаний с указанием их значимости. На их основе осуществляется поиск наиболее релевантных документов относительно заданной тематической эталонной коллекции. Метод позволяет анализировать коллекцию текстов без начального указания количества тематик, что является существенным для создаваемого нами решения. Подход состоит из следующих этапов:

1) Фильтрация стоп-слов. Здесь собраны наиболее часто употребляемые слова (союзы, предлоги, местоимения и др).

2) Расчет характеристик и весов слов\словосочетаний в документе на базе вероятностно-энтропийного метода, основанного на дивергенции Кульбака–Лейблера, энтропийных признаках, модели случайного распределения Бернулли и др. [14].

3) Кластеризация с использованием аффинного преобразования. Из ключевых биграмм строится матрица смежности, описывающая связи между словами в рамках предложения. К этой матрице применяется метод кластеризации на основе близости узлов через соседей (метод аффинного преобразования – Affinity Propagation [15], используя библиотеку scikit-learn [16]), результатом работы алгоритма является набор тематических кластеров, представленных взвешенными ключевыми словами и словосочетаниями.

2.4. Аспектно-тональный анализ

Для разработки метода оценки тональности в составе методики за основу взята архитектура модели нейросети IAN, которая показала хорошие результаты в задаче аспектно-тонального анализа текстов на английском языке [17]. Архитектура модели для определения тональности аспектов текста, состоит из 2 частей, формирующих отдельные представления для целевого объекта (аспекта) и его контекста в интерактивном режиме. Одна из частей сети обрабатывает контекст для рассматриваемого аспекта, в другой обрабатываются слова самого аспекта.

В нашей модели аспектом являются слова принадлежащие одной именованной сущности, для которой определяется тональность (1), а контекст содержит в себе слова предложения, содержащего рассматриваемую именованную сущность (2).

$$target = [w_t^1, w_t^2, \ldots, w_t^N], \quad (1)$$

$$context = [w_c^1, w_c^2, \ldots, w_c^M], \quad (2)$$

где N – количество слов целевого аспекта, M – количество слов в предложении, содержащем целевой аспект.

Извлечение именованных сущностей производится с помощью модели нейронной сети из библиотеки DeepPavlov[3], основанной на топологии BERT [18]. Модель предварительно обучалась на русскоязычной части Википедии и новостных данных [19] с дальнейшим обучением на корпусе Collection3 [20]. Точность модели по метрике f1-score составляет 0.98%, что является одним из state-of-the-art решений для русского языка. В данной работе используются следующие категории сущностей: Персоны, Организации, Локации.

Слова выделенных контекста и аспекта преобразуются в вектора с использованием языковой модели ELMo [6], обученной на корпусе Википедии. Полученные векторные представления слов аспекта и контекста подаются в рекуррентную нейронную сеть на базе LSTM слоев [21] для получения соответствующих скрытых состояний слов. После этого, их средние значения используются для генерации векторов внимания. Далее внутренние представления целевого объекта и контекста объединяются и полученный вектор подается на полносвязный слой с активационной функцией softmax для итогового класса тональности. Схема архитектуры предложенной модели представлена на рис. 2.

Благодаря такой реализации механизма внимания, целевой объект и контекст могут влиять на формирование своих внутренних представлений в интерактивном режиме.

Разработанный метод показал лучшие точности по сравнению с другими решениями при тестировании в задаче аспектно-тонального анализа на корпусе, собранном для соревнования SentiRuEval-2015 [22] (см. табл. 1).

Проведенные эксперименты показали, что использование модели векторного представления слов на базе ELMo для векторизации аспекта и контекста в рамках нейросетевой архитектуры типа IAN позволяет получать более высокие точ-

[3] Модель “ner_rus_bert”: http://docs.deeppavlov.ai/en/master/features/models/ner.html

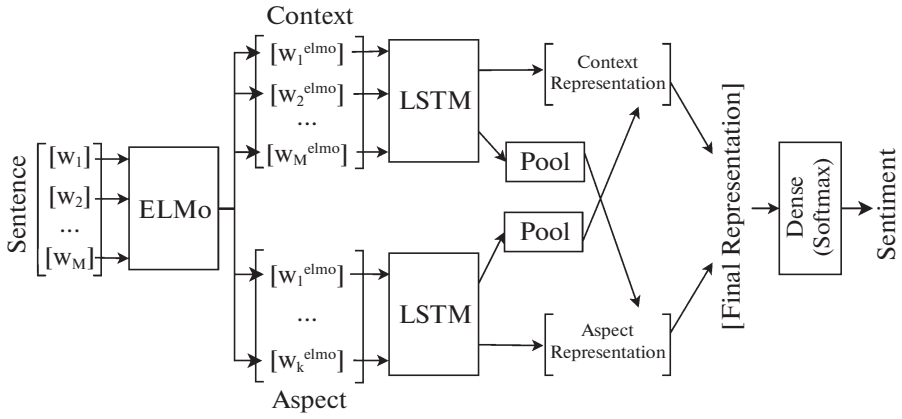

Рис. 2. Архитектура предложенной модели на базе IAN.

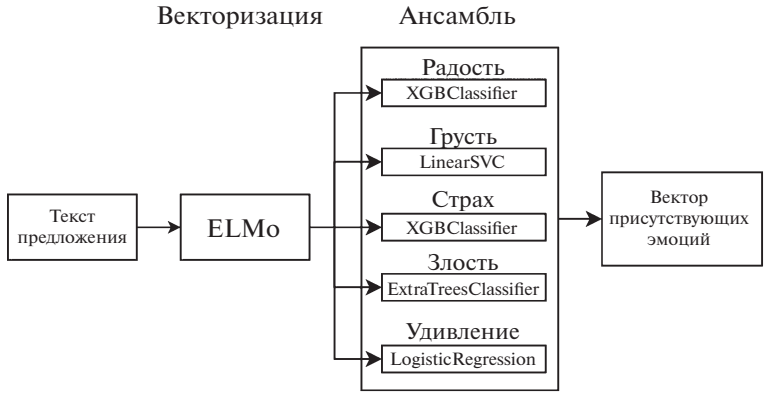

Рис. 3. Архитектура разработанного ансамблевого классификатора эмоций.

ности в задаче аспектной тональности по сравнению с другими реализациями.

2.5. Эмотивный анализ

Для эмотивного анализа на русском языке отсутствуют готовые алгоритмы и размеченные наборы данных, поэтому для реализации функционала эмотивного анализа создан метод на базе контекстно-зависимых векторных представлений слов и последующей обработкой ансамблевым классификатором, обученным на специально подготовленном корпусе текстовых данных для 5 базовых эмоций (радость, грусть, злость, страх и удивление). Корпус состоит из 7235 предложений из постов социальной сети "Живой Журнал" [24], текстов новостного интернет-издания "Лента.Ру"[4] и сообщений микроблога "Твиттер" [25]. Для разметки использовалась краудсорсинговая платформа, где аннотаторам предлагалось оценить предварительно отобранные предложения, которые включают слова из словарей: а) оценочных слов и выражений русского языка РуСентиЛекс [26] и б) эмотивной лексики русского языка[5].

Статистика собранного корпуса представлена в табл. 2.

Для реализации классификатора эмоций нами проведены сравнительные эксперименты методов машинного обучения, среди которых лучшую точность продемонстрировал ансамблевый классификатор на базе признаков, полученных из модели векторного представления слов ELMo. В состав ансамбля входят несколько моделей[6], в том числе: машина опорных векторов с линейным ядром, градиентный бустинг и решающие деревья (см. рис. 3).

При обучении на 80% корпуса и тестировании на 20% разработанный метод показал точность

[4] Корпус текстов издания Лента.Ру: https://github.com/yutkin/Lenta.Ru-News-Dataset

[5] Словарь эмотивной лексики русского языка: https://lexrus.ru/default.aspx?p=2876

[6] Для расчета моделей и метрик используется библиотека scikit-learn: https://scikit-learn.org/

Таблица 1. Полученные точности решения задачи аспектно-тонального анализа

	Отзывы по автомобилям		Отзывы по ресторанам	
	f1-micro	f1-macro	f1-micro	f1-macro
SentiRuEval-2015 – базовая модель [22]	0.62	0.26	0.71	0.27
SentiRuEval-2015 – лучшая модель [23]	0.74	0.57	0.82	0.55
Предложенная модель	**0.79**	**0.60**	**0.85**	**0.58**

Таблица 2. Статистика собранного корпуса текстов по 5 эмоциям

	Радость	Грусть	Страх	Злость	Удивление	Всего
Живой Журнал	316	222	208	174	323	1243
Лента. Ру	153	65	106	102	166	592
Твиттер	1010	1114	251	156	175	2706
Всего	1479	1401	565	432	664	4541

Таблица 3. Результаты сравнительных экспериментов для классификации эмоций (f1-macro)

	Радость	Грусть	Страх	Злость	Удивление	Среднее
Случайный выбор метки	0.45	0.45	0.4	0.38	0.39	0.41
Машина опорных векторов (SVM) на признаках TF-IDF по *n*-граммам символов (от 4 до 8)	0.67	0.69	0.64	0.52	0.68	0.64
Словари эмотивной лексики	0.68	0.64	0.68	0.65	0.67	0.66
Наш подход	**0.88**	**0.83**	**0.77**	**0.67**	**0.74**	**0.78**

выше на 0.37, 0.14 и 0.12 для случайного выбора, базового классификатора на SVM и словарного метода, соответственно (см. табл. 3).

3. ПРИМЕР ВИЗУАЛИЗАЦИИ РЕЗУЛЬТАТОВ АНАЛИЗА

В этой главе приводится пример использования сформированного комплекса инструментов для анализа текстовых данных. Для экспериментов по выделению тематических кластеров были выбраны 2 источника: социальные сети и новостные ресурсы. В качестве данных из социальных сетей был взят имеющийся корпус постов из “Живого Журнала”, а в качестве новостного источника был рассмотрен русскоязычный корпус текстов SCTM-ru [27] с открытой лицензией, в котором использованы статьи международного новостного сайта “Русские Викиновости”. Корпус содержит в себе 7 тыс. документов, 185 авторов и почти 12 тыс. уникальных тематических категорий.

Чтобы в заданных источниках выбрать похожие документы и выделить в них тематики, для задания изначальной анализируемой темы, из “Живого Журнала” была выбрана эталонная коллекция из 10 документов со схожей темой “выборы в Москве”. После этого выделялись ключевые слова и словосочетания, которые использовались для поиска похожих документов в заданных источниках. Документ считался похожим, если в нем присутствовало больше 2 ключевых словосочетаний и 15 ключевых слов. В итоговую коллекцию попало 47 документов из новостного источника и 40 из социальных сетей.

В найденных документах снова выделялись ключевые слова и словосочетания, а также определялись тематики. Автоматически выделено 49 наиболее представленных тематик, а в топ 5 наиболее значимых ключевых слов вошли: “*кандидат*”, “*выборы*”, “*Россия*”, “*партия*”, “*избиратель*”.

Для представления результатов выделения тематических кластеров проводится визуализация в виде графа, узлами которого являются выделенные ключевые слова, а связи отражают встречаемость ключевых слов в ключевой биграмме. Размер узлов зависит от рассчитанного веса ключевого слова. Чем вес слова больше, тем оно важнее для тематики, а узел крупнее. Слова из одной тематики имеют один и тот же цвет. Толщина связи между словами зависит от частоты встречаемости рассматриваемого словосочетания в рассматриваемой коллекции документов. Цвета отражают принадлежность к тематическим кластерам.

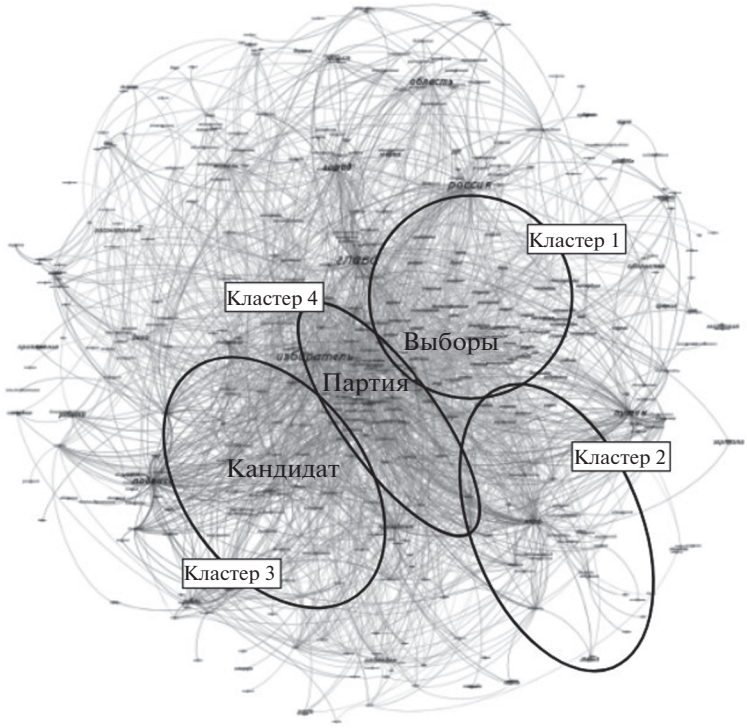

Рис. 4. Граф кластеров (овалы) самых представительных тематик.

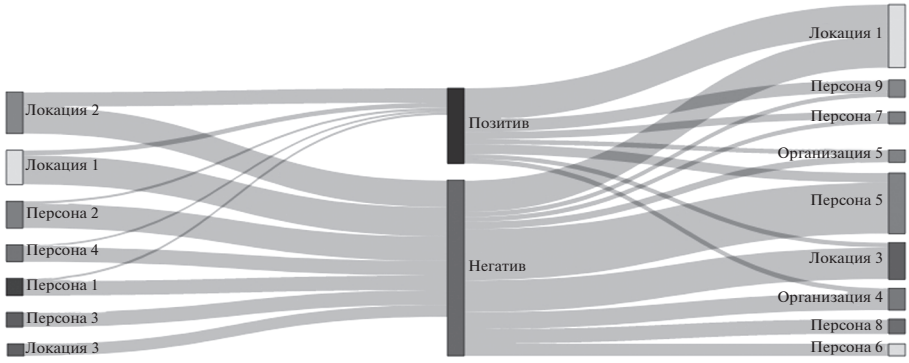

Рис. 5. Диаграмма распределения упоминаний Персон, Организаций и Локаций в тональных контекстах для социальных сетей (слева) и новостей (справа).

Графическое представление графа реализовано с использованием графического редактора Gephi [28]. Для укладки графа используется алгоритм ForceAtlas 2 [29]. В качестве настраиваемых параметров используется коэффициент отталкивания между узлами, со значением 5, гравитация, которая не дает узлам сильно разлетаться, со значением 2, а также включенный запрет перекрытия, чтобы узлы не загораживали друг друга. Пример визуализированного графа выделенных тематик представлен на рис. 4. Далее для выбранной коллекции документов был проведен эмотивно-тональный анализ с выделением именованных сущностей. В топ 5 наиболее часто встречаемых ключевых слов и именованных сущностей среди новостей вошли: *“выборы”, “Россия”, “партия”, “голос”, “кандидат”*; а среди социальных сетей: “*кандидат*”, “*выборы*”, “*партия*”, *“избиратель”, “депутат”*.

Визуализация результатов аспектно-тонального анализа для выделенных именованных сущностей, наиболее часто упоминаемых в позитивном и негативном ключе, представлена на рис. 5.

Из рисунка видно, что в новостных текстах количество упоминаний в разных коннотациях распределено более равномерно, в то время как в со-

Рис. 6. Тематические кластеры для предложений в позитивной тональности из новостного источника (слева) и из социальной сети (справа).

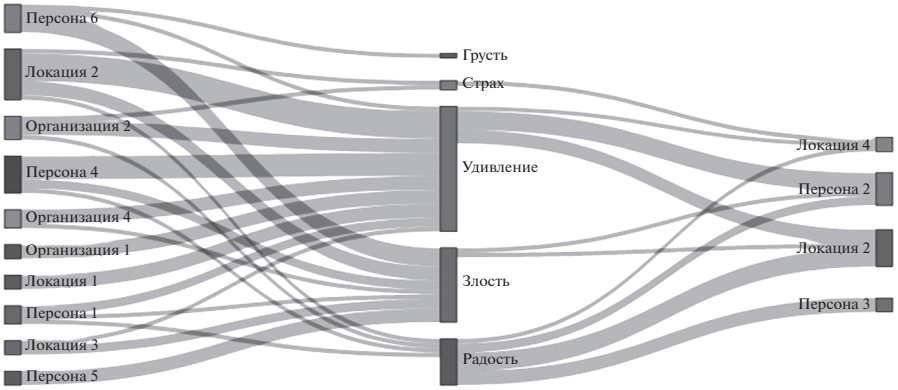

Рис. 7. Диаграмма распределения упоминаний Персон, Организаций и Локаций в различных эмоциях для социальных сетей (слева) и новостей (справа).

циальных сетях контекст зачастую более негативный.

На рис. 6 представлены аспектно-тональные семантические графы тематик для позитивного контекста, выделенных именованных сущностей, из новостей и социальных сетей, соответственно. При этом слова документа, принадлежащие одной именованной сущности, рассматриваются при тематическом анализе как одно слово. Например, слова сущности “Иван Иванович” анализируются как одно слово “Иван_Иванович”.

Визуализация результатов анализа эмотивности выбранных данных (см. рис. 7) позволяет сравнить новостные тексты, где авторы стараются не закладывать эмоции страха и злости в тексты своих публикаций, транслировать в основном позитивную повестку или информировать о необычных событиях, с текстами в социальных сетях, где авторы часто пишут со злостью или негодованием.

Таким образом, предложенный набор инструментов позволяет проводить анализ массива текстовых данных с определением тематик и эмотивно-тональных оценок.

4. ВЫВОДЫ

На базе проведенного исследования разработан инструмент комплексного тематического, аспектно-тонального и эмотивного анализа русскоязычных текстовых данных по апробации и разработке методов интеллектуального анализа данных. В состав комплекса включены разработанные и адаптированные средства для тематической кластеризации, аспектно-тонального анализа и выделения эмоций из авторского текста. Отличительной особенностью разработанного комплекса является возможность проводить автоматизированный тонально-эмотивный анализ текстов с выделением представленных в них тематик. Апробирование комплекса продемонстрировало его эффективность по рассматриваемому набору видов анализа.

5. ФИНАНСИРОВАНИЕ

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-37-00398 и с использованием оборудования центра коллективного пользования "Комплекс моделирования и обработки данных исследовательских установок мега-класса" НИЦ "Курчатовский институт", http://ckp.nrcki.ru/.

СПИСОК ЛИТЕРАТУРЫ

1. IBM Watson Explorer. Available at: https://www.ibm.com/products/watson-explorer (accessed: 14.03.2020).
2. iFORA. Available at: https://issek.hse.ru/news/254274661.html (accessed: 14.03.2020).
3. Semantic Archive Platform. Available at: http://www.anbr.ru/ (accessed: 14.03.2020).
4. *Ma D. et al.* Interactive attention networks for aspect-level sentiment classification. *arXiv preprint* arXiv:1709.00893. 2017.
5. *Huang B., Ou Y., Carley K.M.* Aspect level sentiment classification with attention-over-attention neural networks. *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer, Cham,* 2018. P. 197–206.
6. *Peters M. E. et al.* Deep contextualized word representations. *arXiv preprint* arXiv:1802.05365. 2018.
7. *Pennington J., Socher R., Manning C.* Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014. С. 1532–1543.
8. *Duppada V., Jain R., Hiray S.* Seernet at semeval-2018 task 1: Domain adaptation for affect in tweets. *arXiv preprint* arXiv:1804.06137. 2018.
9. *Jabreel M., Moreno A.* EiTAKA at SemEval-2018 Task 1: An ensemble of n-channels ConvNet and XGboost regressors for emotion analysis of tweets. *arXiv preprint* arXiv:1802.09233. 2018.
10. *Mohammad S., Kiritchenko S.* Understanding emotions: A dataset of tweets to study interactions between affect categories. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.
11. *Mohammad S.M., Bravo-Marquez F.* Emotion intensities in tweets. *arXiv preprint* arXiv:1708.03696. 2017.
12. *Straka M., Straková J.* Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.* 2017. P. 88–99.
13. *Moloshnikov I.A. et al.* A probabilistic-entropy approach of finding thematically similar documents with creating context-semantic graph for investigating evolution of society opinion. *Journal of Physics: Conference Series. IOP Publishing*, 2016. V. 681. № 1. P. 012012.
14. *Moloshnikov I.A., Sboev A.G., Rybka R.B., & Gydovskikh D.V.* An algorithm of finding thematically similar documents with creating context-semantic graph based on probabilistic-entropy approach. *Procedia Computer Science*, 2015. № 66. P. 297–306.
15. *Frey B.J., Dueck D.* Clustering by passing messages between data points. *Science.* 2007. V. 315. № 5814. P. 972–976.
16. *Pedregosa F., Varoquaux и др.*, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*. 2011. № 12. С. 2825–2830.
17. *Ma D. et al.* Interactive attention networks for aspect-level sentiment classification. *arXiv preprint* arXiv:1709.00893. 2017.
18. *Devlin J. et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805. 2018.
19. *Kuratov Y., Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *arXiv preprint* arXiv:1905.07213. 2019.
20. *Mozharova V., Loukachevitch N.* Two-stage approach in Russian named entity recognition. *International FRUCT Conference on Intelligence, Social Media and Web, ISMW FRUCT 2016.* Saint-Petersburg; Russian Federation, https://doi.org/10.1109/FRUCT.2016.7584769
21. *Hochreiter S., Schmidhuber J.* Long short-term memory. *Neural computation*. 1997. Т. 9. № 8. С. 1735–1780.
22. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., & Tutubalina E.* SentiRuEval: testing object-oriented sentiment analysis systems in Russian. In *Proceedings of International Conference Dialog.* 2015. V. 2. P. 3–13.
23. *Blinov P., Kotelnikov E.V.* Semantic similarity for aspect-based sentiment analysis. *Russian Digital Libraries Journal.* 2015. Т. 18. № 3–4. С. 120–137.
24. Rusprofiling corpus of russian texts. Available at http://rusprofilinglab.ru/rusprofiling-atpan/corpus/ (accessed: 14.03.2020)
25. *Rubtsova Y.* Avtomaticheskoe postroenie i analiz korpusa korotkih tekstov (postov mikroblogov) dlja zadachi razrabotki i trenirovki tonovogo klassifikatora [Automatic construction and analysis of the short texts dataset (microblogging posts) for the task of developing and training sentiment classifier]. *Inzhenerija znanij i tehnologii semanticheskogo veba*. 2012. Т. 1. С. 109–116.
26. *Loukachevitch N., Levchik A.* Creating a general russian sentiment lexicon. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016.
27. *Karpovich S.N.* The Russian language text corpus for testing algorithms of topic model. *SPIIRAS Proceedings.* 2015. Т. 39. С. 123–142.
28. *Bastian M., Heymann S., Jacomy M.* Gephi: an open source software for exploring and manipulating networks. *Third international AAAI conference on weblogs and social media.* 2009.
29. *Jacomy M. et al.* ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one.* 2014. № 6. Т. 9. P. e98679.

Vestnik Natsional'nogo issledovatel'skogo yadernogo universiteta "MIFI", 2020, vol. 9, no. 3, pp. 279–288

Method for Automated Intelligent Emotive and Sentiment Analysis of Texts with a Thematic Focus

A. V. Naumov[a,#], A. A. Selivanov[a,##], I. A. Moloshnikov[a,###], and A. G. Sboev[a,b,####]

[a] *National Research Center Kurchatov Institute, Moscow, 123182 Russia*

[b] *National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, 115409 Russia*

[#]*e-mail: sanya.naumov@gmail.com, dmitrygagus@gmail.com*

[##]*e-mail: aaselivanov.10.03@gmail.com*

[###]*e-mail: ivan-rus@yandex.ru*

[####]*e-mail: sag111@mail.ru*

Received June 23, 2020; revised June 30, 2020; accepted September 8, 2020

Abstract—A tool for complex thematic, tonal, aspect-sentiment, and emotive analysis of natural-language texts is reported. For thematic analysis, a method based on probabilistic-entropy characteristics is employed, which allows us to identify automatically the number of topics in a text collection. Aspect-based sentiment analysis is performed within the neural network model with the topology of an interactive attention network, which demonstrates an accuracy of 0.58 by F1-macro score on the corpus from the SentiRuEval-2015 competition, thus outperforming the best results of the competition. For conducting emotive analysis, a method is developed on the basis of context-dependent vector representations of words, with the subsequent processing by an ensemble classifier trained on a corpus of texts prepared specially for five basic emotions: joy, sadness, anger, fear, and surprise (this classifier achieving an accuracy of 0.76 by F1-macro). An example of using the method is also demonstrated. Texts of the LiveJournal social network and news from the SCTM-ru project are selected for analysis. Presented visualization of text analysis results in the form of graphs, which shows the efficiency of the developed method.

Keywords: data analysis, text data, natural language processing, aspect-based sentiment analysis, emotion recognition, neural networks, machine learning

DOI: 10.1134/S2304487X20030086

REFERENCES

1. IBM Watson Explorer. Available at: https://www.ibm.com/products/watson-explorer (accessed: 14.03.2020).
2. iFORA. Available at: https://issek.hse.ru/news/254274661.html (accessed: 14.03.2020).
3. *Semantic Archive Platform*. Available at: http://www.anbr.ru/ (accessed: 14.03.2020).
4. Ma D. et al. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint* arXiv:1709.00893. 2017.
5. Huang B., Ou Y., Carley K.M. Aspect level sentiment classification with attention-over-attention neural networks. *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer, Cham,* 2018. pp. 197–206.
6. Peters M.E. et al. Deep contextualized word representations. *arXiv preprint* arXiv:1802.05365. 2018.
7. Pennington J., Socher R., Manning C. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014. C. 1532–1543.
8. Duppada V., Jain R., Hiray S. Seernet at semeval-2018 task 1: Domain adaptation for affect in tweets. *arXiv preprint* arXiv:1804.06137. 2018.
9. Jabreel M., Moreno A. EiTAKA at SemEval-2018 Task 1: An ensemble of n-channels ConvNet and XGboost regressors for emotion analysis of tweets. *arXiv preprint* arXiv:1802.09233. 2018.
10. Mohammad S., Kiritchenko S. Understanding emotions: A dataset of tweets to study interactions between affect categories. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.
11. Mohammad S.M., Bravo-Marquez F. Emotion intensities in tweets. *arXiv preprint* arXiv:1708.03696. 2017.
12. Straka M., Straková J. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. 2017. pp. 88–99.
13. Moloshnikov I.A. et al. A probabilistic-entropy approach of finding thematically similar documents with creating context-semantic graph for investigating evo-

lution of society opinion. *Journal of Physics: Conference Series. IOP Publishing*, 2016. № 1. Т. 681. p. 012012.

14. Moloshnikov I.A., Sboev A.G., Rybka R.B., & Gydovskikh D.V.. An algorithm of finding thematically similar documents with creating context-semantic graph based on probabilistic-entropy approach. *Procedia Computer Science*, 2015. №66. pp. 297–306.
15. Frey B.J., Dueck D. Clustering by passing messages between data points. *Science.* 2007. № 5814. vol. 315. pp. 972–976.
16. Pedregosa F. and Varoquaux и др., "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, 2011, № 12, pp. 2825–2830.
17. Ma D. et al. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint* arXiv:1709.00893. 2017.
18. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805. 2018.
19. Kuratov Y., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *arXiv preprint* arXiv:1905.07213. 2019.
20. Mozharova V., Loukachevitch N., Two-stage approach in Russian named entity recognition. *International FRUCT Conference on Intelligence, Social Media and Web, ISMW FRUCT 2016.* Saint-Petersburg; Russian Federation, DOI 10.1109/FRUCT.2016.7584769
21. Hochreiter S., Schmidhuber J. Long short-term memory. *Neural computation*. 1997. Т. 9. № 8. С. 1735–1780.
22. Loukachevitch, N., Blinov, P., Kotelnikov, E., Rubtsova, Y., Ivanov, V., & Tutubalina, E. SentiRuEval: testing object-oriented sentiment analysis systems in Russian. In *Proceedings of International Conference Dialog.* 2015. Vol. 2, pp. 3–13.
23. Blinov P., Kotelnikov E.V. Semantic similarity for aspect-based sentiment analysis. *Russian Digital Libraries Journal.* 2015. Т. 18. № 3–4. С. 120–137.
24. Rusprofiling corpus of russian texts. Available at http://rusprofilinglab.ru/rusprofiling-atpan/corpus/ (accessed: 14.03.2020)
25. Rubtsova Y. Avtomaticheskoe postroenie i analiz korpusa korotkih tekstov (postov mikroblogov) dlja zadachi razrabotki i trenirovki tonovogo klassifikatora [Automatic construction and analysis of the short texts dataset (microblogging posts) for the task of developing and training sentiment classifier]. *Inzhenerija znanij i tehnologii semanticheskogo veba*. 2012. Т. 1. С. 109–116.
26. Loukachevitch N., Levchik A. Creating a general russian sentiment lexicon. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016.
27. Karpovich S.N. The Russian language text corpus for testing algorithms of topic model. *SPIIRAS Proceedings.* 2015. Т. 39. С. 123–142.
28. Bastian M., Heymann S., Jacomy M. Gephi: an open source software for exploring and manipulating networks. *Third international AAAI conference on weblogs and social media*. 2009.
29. Jacomy M. et al. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PloS one. 2014. № 6. Т. 9. p. e98679.