УДК 004.912 **DOI** 10.15622/ia.23.4.9

А.А. Двойникова, И.А. Кагиров, А.А. Карпов

МЕТОД РАСПОЗНАВАНИЯ СЕНТИМЕНТА И ЭМОЦИЙ В ТРАНСКРИПЦИЯХ РУССКОЯЗЫЧНОЙ РЕЧИ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ПЕРЕВОДА

Двойникова А.А., Кагиров И.А., Карпов А.А. **Метод распознавания сентимента и эмоций в транскрипциях русскоязычной речи с использованием машинного перевода.**

Аннотация. В статье рассматривается проблема распознавания сентимента и эмоций пользователей в русскоязычных текстовых транскрипциях речи с использованием словарных методов и машинного перевода. Количество имеющихся информационных ресурсов для анализа сентимента текстовых сообщений на русском языке очень ограничено, что существенно затрудняет применение базовых методов анализа сентимента, а именно, предобработки текстов, векторизации с помощью тональных словарей, традиционных классификаторов. Для решения этой проблемы в статье вводится новый метод на основе автоматического машинного перевода русскоязычных текстов на английский язык. Частичный перевод предполагает перевод отдельных лексем, не включенных в русскоязычные тональные словари, тогда как полный перевод подразумевает перевод всего текста целиком. Переведенный текст анализируется с использованием различных англоязычных тональных словарей. Экспериментальные исследования для решения задачи распознавания сентимента и эмоций были проведены на текстовых транскрипциях многомодального русскоязычного корпуса RAMAS, извлеченных из аудиоданных экспертным путем и автоматически с использованием системы распознавания речи. В результате применения методов машинного перевода достигается значение взвешенной F-меры распознавания семи классов эмоций 31,12 % и 23,74 %, и трех классов сентимента 75,37 % и 71,60 % для экспертных и автоматических транскрипций русскоязычной речи корпуса RAMAS, соответственно. Также в ходе экспериментов было выявлено, что использование статистических векторов в качестве метода преобразования текстовых данных позволяет достичь значение показателя взвешенной F-меры на 1-5 % выше по сравнению с использованием конкатенированного (статистического и тонального) вектора. Таким образом, эксперименты показывают, что объединение всех англоязычных тональных словарей позволяет повысить точность распознавания сентимента и эмоций в текстовых данных. В статье также исследуется корреляция между длиной вектора текстовых данных и его репрезентативностью. По результатам экспериментов можно сделать вывод, что использование лемматизации для нормализации слов текстовых транскрипций речи позволяет достичь большей точности распознавания сентимента по сравнению со стеммингом. Использование предложенных методов с полным и частичным машинным переводом позволяет повысить точность распознавания сентимента и эмоций на 0,65–9,76 % по показателю взвешенной F-меры по сравнению с базовым методом распознавания сентимента и эмоций.

Ключевые слова: машинный перевод, тональные словари, распознавание эмоций, сентимент-анализ, тональные вектора.

1. Введение. Анализ тональности текста (сентимент-анализ, англ. sentiment analysis) – это область компьютерной лингвистики, связанная с методами определения эмоциональной полярности текста

на естественном языке. Анализ тональности является частным случаем извлечения информации (англ. data mining), однако он не подразумевает извлечения имен сущностей, ограничиваясь только эмоциональной окраской текстов [1].

Анализ тональности текстовых сообщений актуален во многих сферах человеческой деятельности: оценка качества товаров и услуг, мониторинг общественного мнения, прогнозы на основе новостных подборок в Интернете и т.п. [2 – 3]. Другой важной областью является межчеловеческое взаимодействие в виртуальном пространстве, подразумевающее коммуникацию на естественном языке [4]. Благодаря важности перечисленных областей, анализ сентимента является динамичной и быстро развивающейся отраслью компьютерной лингвистики и методов анализа естественного языка в целом.

В общем случае анализ сентимента сводится к отнесению конкретного текста или группы текстов к определенному классу в зависимости от эмоциональной валентности текста. В существующих исследованиях используют различные классификации сентимента: бинарная (негативный, позитивный), тернарная (негативный, нейтральный, позитивный) и многоуровневую (от сильно негативного до сильно позитивного) [5].

Стоит отметить смежную область – автоматический анализ эмоций в тексте. В классической работе [6], теоретические результаты из которой фактически заложили основы современных исследований в области распознавания эмоций, было выделено шесть базовых эмоций, и для их нахождения используется две группы методов: основанные на словарях и основанные на корпусах. Первый метод достаточно прямолинеен и напрямую зависит от доступных словарей эмоций для конкретного языка. Второй метод подразумевает построение математической модели на основе текстов, предварительно размеченных экспертами. Несмотря на наличие определенной корреляции между определением сентиментом и эмоций, следует иметь в виду, что эмоции – это выражение психофизиологических состояний индивида, а сентимент – отношение говорящего к определенной теме [7].

Настоящая статья посвящена улучшению методов автоматического анализа сентимента и эмоций русскоязычных транскрипций речи за счет использования машинного перевода текстовых данных. Количество информационных ресурсов (корпусов данных, тональных словарей) для анализа текстов на русском языке на сегодня остается достаточно ограниченным

[8 – 9], в связи с чем предлагается метод, основанный на применении автоматического машинного перевода русскоязычных текстов на английский язык.

Автоматический машинный перевод текстовых данных используется в существующих исследованиях для распознавания сентимента [10 – 11] и эмоций [12]. Такой метод актуален для малоресурсных языков, потому что из текстов на таких языках достаточно сложно извлечь лингвистические признаки. Другой причиной может являться маленький объем данных. Поэтому машинный перевод используется в прямом виде (для извлечения лингвистических признаков на другом языке или увеличении объема тональных словарей) [11, 13] или как двойной обратный перевод (для аугментации обучающего набора данных).

Часто производится двойной обратный перевод: с исходного языка на другой выбранный, затем с выбранного обратно на исходный. В большинстве случаев в качестве промежуточного языка выбирается английский язык [10]. Метод машинного перевода для улучшения точности классификации сентимента или эмоций используется для менее ресурсных языков, например для турецкого [10], словацкого [11], иврита [14], польского [12], испанского [14], русского [15] и других.

2. Исследовательский корпус речевых данных. Для экспериментальных исследований использован русскоязычный многомодальный корпус данных RAMAS [16]. Он содержит 581 аудио- и видеозапись с участием десяти актеров общей продолжительностью 395 минут. Особенностью корпуса является то, что актеры попарно разыгрывали диалоги по диадическим сценариям. Сценарии были составлены таким образом, чтобы каждый диктор проявил одну из шести основных эмоций: радость, страх, удивление, гнев, грусть, отвращение. Также в корпусе присутствуют монологи каждого диктора с эмоционально нейтральной речью. Для распознавания сентимента и эмоций в качестве меток текста использовались значения эмоций, указанные в сценариях.

Корпус RAMAS размечен только по классам эмоций, разметка по сентиментам высказываний не проводилась. Для построения системы распознавания сентимента необходимо сгруппировать классы эмоций по принципу валентности эмоций. Обоснованностью такого преобразования с точки зрения психологии являются работы американского психолога Дж. Рассела. Диаграмма эмоций Рассела [17] – модель, созданная для описания и классификации эмоций на основе двух основных измерений: валентности и активации. Валентность

относится к тому, насколько положительно или отрицательно оценивается эмоция. Активация в диаграмме Рассела относится к уровню возбуждения психики человека, связанному с эмоцией. Основываясь на диаграмме Рассела в настоящем исследовании для выделения групп сентимента такие эмоции как радость, удивление группируется в положительный класс сентимента, страх, грусть, гнев и отвращение – негативный класс, а нейтральное состояние относится к нейтральному классу. Таким образом, все высказывания корпуса RAMAS группируются в три класса сентимента.

Следует отметить, что материал, содержащийся в корпусе данных, содержит шумы, речь одновременно нескольких дикторов или организаторов. Кроме того, авторы RAMAS не предоставили расшифровок (орфографических транскрипций) речи дикторов. Поэтому для экспериментальных исследований из аудиозаписей диалогов были извлечены транскрипции реплик дикторов экспертным (человеком-аудитором) и автоматическим методом. Для автоматического распознавания речи (APP) использовались сервисы SpeechKit [1] от компании Яндекс и SpeechRecognition [2] от компании Google. Итоговый текстовый корпус составил 535 транскрипций произнесенных текстов, извлеченных экспертным методом, и 263 текста с использованием систем APP. Разница в количестве извлеченных текстов (535 и 263) обусловлена особенностями данных, содержащихся в корпусе RAMAS, а именно, большим количеством шумов и наложений речи нескольких дикторов. Поэтому от большого числа транскрипций, полученных автоматическим методом, пришлось отказаться. Также стоит отметить, что автоматические транскрипции содержат в себе грамматические и лексические ошибки. В ходе эксперимента оба набора транскрипций – как экспертные, так и автоматические – были отдельно использованы для проверки предложенных гипотез.

На рисунке 1 представлено распределение текстов в каждом классе сентимента и эмоций для экспертных и автоматических транскрипций.

[1] https://cloud.yandex.ru/ru/services/speechkit
[2] https://cloud.google.com/speech-to-text/

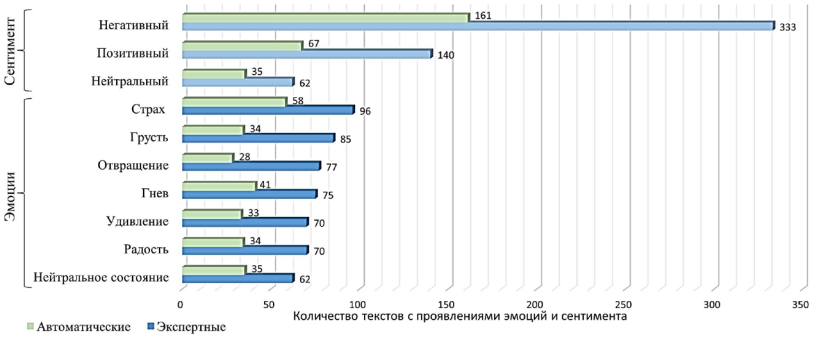

Рис. 1. Гистограмма количества текстов с проявлениями сентимента и эмоций в экспертных (синий) и автоматических (зеленый) транскрипциях корпуса RAMAS

В результате анализа диалогов, содержащихся в базе данных RAMAS, было выявлено количественное соотношение трех видов сентимента – позитивной, негативной и нейтральной, с преобладанием негативного сентимента как для экспертных транскрипций, так и для автоматических. Как можно заметить из рисунка 1 распределение сентиментов в данных сильно несбалансировано, при этом количество текстов в классах эмоций почти одинаковое.

Для анализа экспертных и автоматических транскрипций корпуса RAMAS были построены диаграммы размаха длин текстовых данных, они представлены на рисунке 2.

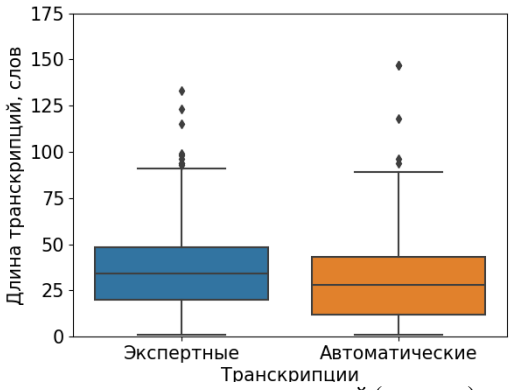

Рис. 2. Диаграммы размаха длин транскрипций (в словах) из корпуса RAMAS

Из рисунка 2 видно, что максимальное значение количества слов в высказываниях (за исключением выбросов) достигает 90. При этом медианное значение длин всех текстов находится в значении около 30-35 слов.

3. Метод распознавания сентимента и эмоций. Для распознавания сентимента и эмоций пользователей предложен метод, представленный в виде блок-схемы на рисунке 3.

Рис. 3. Блок-схема метода распознавания сентимента и/или эмоций в русскоязычных текстовых транскрипциях речи

Все серые блоки со сплошным контуром образуют базовый метод. Блоки с пунктирным контуром выполняются опционально. Также в статье исследуются методы машинного перевода для увеличения точности распознавания сентимента и эмоций в русскоязычных текстовых данных, рассматривается возможность использования полного и частичного перевода текстовых данных.

Блок предобработки текстовых данных включает в себя токенизацию (разделение текста на токены – слова), понижение регистра, удаление пунктуации и стоп-слов, а также лемматизацию или стемминг. При лемматизации слово приводится к его начальной форме (лемме) с использованием морфологических правил и словарей, позволяющих получить слово в одной из его форм в контексте и привести к словарной форме. Стемминг, наоборот, сводится к удалению всех аффиксов словоформы, не входящих в основу слова. Для стемминга используются эвристические правила и алгоритмы, которые могут приводить к ошибкам, если правила не учитываются в полной мере.

Для блока векторизации необходимо использование тональных словарей. Под векторизацией понимается преобразование текстовых данных в числовой формат. Векторизация с использованием тональных словарей выполняется по алгоритму, представленному на рисунке 4.

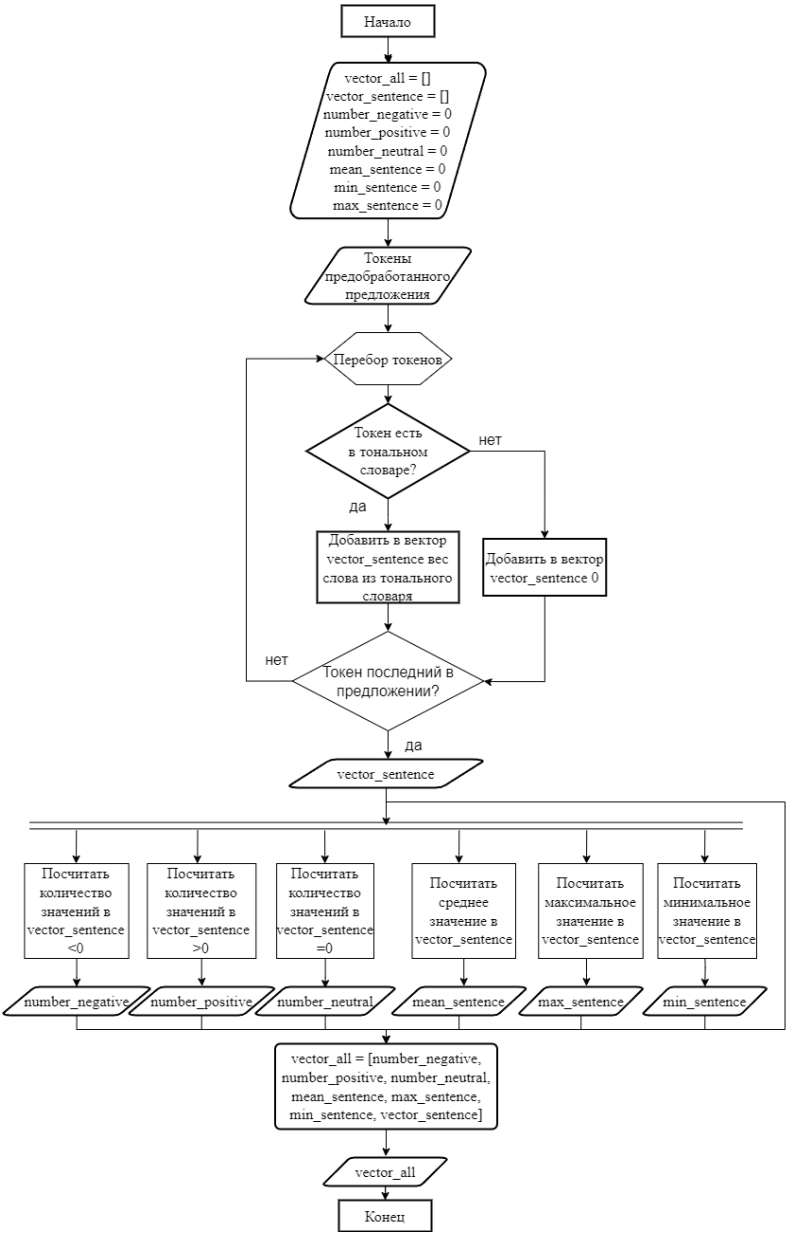

Рис. 4. Блок-схема алгоритма векторизации текстовых предложений с использованием тональных словарей

На рисунке 4 переменная vector_sentence является списком, в котором хранятся значения весов из тонального словаря для каждого токена предложения. Переменные number_negative, number_positive, number_neutral, mean_sentence, max_sentence, min_sentence, sum_pos_sentence и sum_neg_sentence означают количество отрицательных, положительных и нейтральных весов в vector_sentence, а также среднее, максимальное и минимальное значения vector_sentence, сумма весов положительных и отрицательных значений в vector_sentence. Все восемь переменных образуют статистический вектор проявлений сентимента и/или эмоций в тексте. Для распознавания сентимента используются тональные словари, веса слов в которых обозначают степень положительного или отрицательного значения слова, для распознавания эмоций – веса обозначают принадлежность к различным эмоциям.

Предложенный метод векторизации текста имеет ряд недостатков: 1) вектора разных предложений имеют различную длину, 2) значения векторов находятся в разных диапазонах. Для решения первого недостатка все вектора обрезаются или дополняются нулями до длины 98. Выбор такого значения обусловлен максимальным значением длин всех текстов (рисунок 2) в исследуемом корпусе. Затем все вектора нормализуются по следующей формуле (Min-Max нормализация):

$$x_{norm}(i) = \frac{x(i) - \min(x)}{\max(x) - \min(x)}, \quad (1)$$

где $x_{norm}(i)$ и $x(i)$ – нормированное и исходное i значения вектора, $\max(x)$ и $\min(x)$ – максимальное и минимальное значения данного вектора.

Для определения наиболее подходящего классификатора проведено уменьшение размерности данных при помощи линейного дискриминантного анализа (англ. Linear Discriminant Analisys, LDA[3]). Преимущества метода LDA заключается в том, что при преобразовании исходных векторов в пространство необходимой размерности учитываются метки классов. Суть алгоритма LDA заключается в максимизации линейных дискриминант (представленных в виде осей на рисунке 5) каждого класса. Значения линейных дискриминант на осях не имеют информативного характера,

[3]https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html

они необходимы для расчета расстояния Махаланобиса между группами (классами). Значение расстояния Махаланобиса отображает степень разделимости классов. На рисунке 5 изображены вектора исследуемого корпуса данных в двумерном признаковом пространстве.

Как показывает анализ результатов LDA (рисунок 5), достаточно плохо разделяются классы данных как для сентимента, так и для эмоций. В связи с этим в настоящем исследовании для классификации данных использован ядерный метод опорных векторов (англ. Kernel Support Vector Machine, Kernel SVM [4]), являющийся быстрым и точным методом решения задач, связанных с классификацией текстовых данных [5].

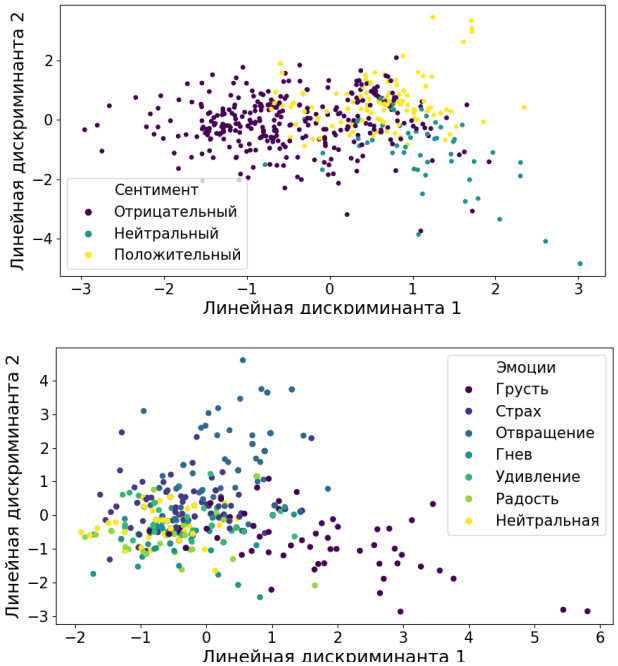

Рис. 5. Распределение векторов в двухмерном пространстве по классам сентимента (сверху) и эмоций (снизу)

Методы машинного перевода. Эмоциональная окраска конкретных языковых единиц, с точки зрения лингвистики,

[4] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC

представляет собой проявление модальности – структурно сложной и многоплановой семантической зоны языка. Согласно [18], модальные значения концентрируются вокруг двух ядерных значений – отношения говорящего к ситуации и статуса самой ситуации по отношению к реальному миру. В предложении, содержащем модальный компонент, всегда заключено отношение говорящего к тому, что он сообщает, иными словами, оценка ситуации.

Эмоциональная окраска текста коррелирует с одним из важнейших типов модальности – оценочной модальностью («хорошо – плохо»), или этической оценкой. Характерно, что значения оценочной модальности имеют тенденцию выражаться лексическими средствами [19], а не грамматическими. Таким образом, вероятность того, что одни и те же значения из области оценочной модальности, связанные с этической оценкой ситуации, будут выражаться сходными средствами для большинства языковых пар, достаточно высока, и прямой перевод лексики, маркированной в отношении оценочной модальности («эмоционально окрашенные слова») с высокой долей вероятности позволит расширить существующие тональные словари для русского языка.

Для русского языка существует ограниченное количество тональных словарей, относительно английского языка. Поэтому применение автоматического машинного перевода позволяет решить проблему недостатка информационных ресурсов русского языка и позволит извлекать большее количество информации о проявлениях сентимента и эмоций в русскоязычных текстах. Для машинного перевода использовалась встроенная модель машинного перевода модуля Translation библиотеки googletrans [5] языка Python. В работе проводились исследования по использованию двух видов машинного перевода: частичного и полного.

Частичный перевод подразумевает под собой перевод на английский язык только тех слов в тексте, которых нет в русскоязычном тональном словаре. Такой перевод выполняется после блока предобработки текстов на русском языке для того, чтобы перевод выполнялся корректно – переводились только леммы слов, а не использовались различные морфологические формы слова. Полный перевод выполняется с исходным русскоязычным текстом для того, чтобы сохранить синтаксическую структуру предложений. После перевода текст векторизуется с использованием англоязычных тональных словарей.

[5] https://pypi.org/project/googletrans/

4. Тональные словари. Одним из методов распознавания сентимента и эмоций в текстовых данных является использование словарей оценочной лексики (тональных словарей, словарей эмоциональных слов). Словари оценочной лексики содержат лексические единицы (слова и словосочетания), каждой из которых присвоена некоторая «эмоциональная оценка». Таким образом, каждой единице словаря задается вес принадлежности к эмоциональному классу [20]. Разметка может быть бинарная, тернарная и многоклассовая (содержащая больше трех классов).

Одним из основных недостатков этого метода является сильная зависимость от наличия тональных словарей. В том случае, если информационное обеспечение для конкретного языка (в данном случае – для русского) недостаточно, созданная система автоматического анализа сентимента текста предсказуемо даст неудовлетворительные результаты. Составление новых словарей является трудоемкой задачей, требующей длительного времени и дополнительного привлечения специалистов. Ниже в таблице 1 представлено описание тональных словарей для русского языка, которые использовались в данном исследовании.

Как видно, анализ тональности имеет определенные информационные ресурсы для русского языка. Однако, в отличие от других основных мировых языков (в первую очередь, английского), системы анализа русскоязычных текстов достигают меньшей точности распознавания сентимента по сравнению с англоязычными, что может быть обусловлено меньшим объемом тональных словарей и обучающих баз данных. Так, в работе [21] демонстрируется, что точность анализа сентимента для английского языка на сегодняшний день выше, чем для другого славянского языка – чешского. В таблице 1 также представлено описание существующих англоязычных тональных словарей, которые использовались в настоящем исследовании. Значения сентимента [-1, 1] и [-4, 4] означают регрессивную аннотацию данных по сентименту. Чем больше диапазон значений сентимента, тем больше вариативность меток аннотаций.

Как можно заметить из таблицы 1, количество и объем тональных словарей в свободном доступе для английского языка выше относительно русскоязычных словарей. При этом количество словарей для английского языка, имеющих оценочную лексику по различным классам эмоций, значительно больше, чем для русского языка. Более представительные информационные ресурсы для английского языка могут помочь извлекать большее количество репрезентативных

признаков для анализа сентимента и эмоций в текстовых данных. Регрессивные оценки сентимента присутствуют только в англоязычных словарях.

5. Экспериментальные исследования

5.1. Нормализация слов транскрипций. Для решения задачи распознавания сентимента и эмоций необходимо выполнить предобработку текстовых данных. В зависимости от различных типов данных, языка и пр. стоит подбирать эффективную комбинацию методов предобработки, с помощью которой можно достичь наиболее высокой точности распознавания сентимента и эмоций. Одним из главных методов предобработки является вид нормализации слов в текстах, который может значительно влиять на точность распознавания сентимента и эмоций. Нормализация слов в текстовых данных может выполняться двумя методами: 1) лемматизация и 2) стемминг. Стоит учесть, что при использовании стемминга слов в тексте в качестве предобработки данных и векторизации текста с использованием тональных словарей, необходимо также обрабатывать слова из тональных словарей стеммингом.

Таблица 1. Описание тональных словарей

Название словаря	**Число слов**	**Количество и содержание классов**
Русскоязычные тональные словари		
RuSentiLex[6] [9]	16057	3 сентимента, смешанная оценка
LinisCrowd [7] [22]	7545	Сильно отрицательные, отрицательные, нейтральные, положительные, сильно положительные
WordNetAffect[8] [23]	2401	6 эмоций (радость, страх, гнев, печаль, отвращение, удивление)
Англоязычные тональные словари		
SentiWordNet[9]	206942	Сентимент [-1, 1]
NRC[10] [24]	1515	Сентимент [-1, 1], 8 эмоций
Bing Liu's Opinion Lexicon[11] [25]	6787	7 эмоций, 3 сентимента
Vader[12] [26]	7520	Сентимент [-4, 4]

[6] https://www.labinform.ru/pub/rusentilex/index.htm
[7] http://linis-crowd.org/
[8] http://lilu.fcim.utm.md/resourcesRoRuWNA_ru.html
[9] https://github.com/aesuli/SentiWordNet
[10] https://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
[11] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon
[12] https://github.com/cjhutto/vaderSentiment

Экспериментальные исследования по определению наиболее эффективного (позволяющего достичь наибольшей точности распознавания сентимента и эмоций) метода нормализации слов в русскоязычных транскрипциях корпуса RAMAS представлены в таблице 2. Лемматизация слов выполнялась с помощью инструмента pymorphy2, а стемминг с помощью nltk. Для проведения данных исследований экспертные транскрипции (ЭТ) и автоматические (АТ) транскрипции векторизовались с использованием объединенного тонального словаря, включающего RuSentiLex, LinisCroes и WordNetAffect. Объединение тональных словарей позволяет извлечь из текстов большее количество эмоционально репрезентативной информации [15]. В качестве машинного классификатора использовался Kernel SVM с подбором гиперпараметров для каждого эксперимента. В данном исследовании использовались следующие ядра: линейное, полиномиальное, сигмовидное, ядро радиальной базисной функции. Обучение классификатора происходило с помощью трехблочной кросс-валидации. Разделение данных на три блока происходило стратифицированным способом. Эксперименты проводились для различных задач классификации: распознавание семи классов эмоций – радость, удивление, грусть, страх, гнев, отвращение, нейтральное состояние и трех классов сентимента – негативный, нейтральный, позитивный. В качестве сравнительной метрики оценки эффективности предложенных методов использовалась взвешенная F-мера. Для многоклассовой классификации она вычисляется по формуле (2).

$$wF = \frac{\sum_{i=0}^{m}(F_i * N_i)}{m}, \tag{2}$$

где wF – взвешенная F-мера для всех классов m, F_i – взвешенная F-мера для i класса, N_i – количество экземпляров в i классе.

Таблица 2. Результаты распознавания сентимента и эмоций в зависимости от метода нормализации слов, взвешенная F-мера (%)

Метод нормализации	**Эмоции**		**Сентимент**	
	ЭТ	АТ	ЭТ	АТ
Лемматизация	**27,42**	**23,09**	**65,61**	**67,79**
Стемминг	26,67	21,87	61,47	61,20

Как видно из таблицы 2, лемматизация в качестве метода нормализации слов в текстовых транскрипциях превосходит стемминг на 1–6% по показателю взвешенной F-меры в зависимости от типа транскрипций (экспертные или автоматические) и задачи классификации (сентимент или эмоции). Более высокое качество классификации сентимента и эмоций при использовании лемматизации может быть связано с возникающими проблемами омонимии в тональных словарях при использовании стемминга.

5.2. Машинный перевод. В данном исследовании проверяется гипотеза об эффективности использования машинного перевода для распознавания сентимента и эмоций в текстовых транскрипциях речевых высказываний. Обзор существующих исследований показал, что английский язык имеет большое количество информационно-лингвистических ресурсов, и качество распознавания сентимента в англоязычных текстах выше, по сравнению с русскоязычными. Поэтому для экспериментальных исследований использовался машинный перевод на английский язык. Для того, чтобы определить наиболее репрезентативный англоязычный тональный словарь для распознавания сентимента, был произведен полный автоматический перевод всех транскрипций корпуса RAMAS на английский язык. Векторизация англоязычных текстовых данных была выполнена как при помощи каждого тонального словаря английский тональных слов, описанных в разделе 5, так и объединенном тональном словаре. Результаты экспериментов представлены в таблице 3.

Таблица 3. Результаты распознавания сентимента при использовании полного машинного перевода на английский язык транскрипций корпуса RAMAS, взвешенная F-мера (%)

Тональный словарь	**Экспертные транскрипции**	**Автоматические транскрипции**
SentiWordNet	63,64	61,51
NRC [24]	65,33	57,36
Bing Liu's Opinion Lexicon [25]	67,91	**66,62**
Vader [26]	71,80	66,52
Объединенный	**75,37**	64,77

Как можно заметить, объединенный англоязычный тональный словарь для распознавания сентимента в переведенных на английский язык транскрипциях показывает лучшую точность только при анализе экспертных транскрипций корпуса RAMAS. Наиболее высокая точность распознавания сентимента в переведенных автоматических транскрипциях достигается при использовании во время векторизации

англоязычного тонального словаря Bing Liu's Opinion Lexicon [25]. Это может быть связано с тем, что автоматические транскрипции содержат в себе много зашумленных данных, не относящихся к речи анализируемого диктора. Слишком большое количество распознанных тональных слов различных дикторов может привезти к ошибкам в распознавании сентимента в высказываниях анализируемого диктора. Поэтому в дальнейших экспериментальных исследованиях для анализа экспертных транскрипций будет использоваться объединенный англоязычный тональный словарь, а для автоматических транскрипций – Bing Liu's Opinion Lexicon [25].

В ходе экспериментов были исследованы и сравнены два типа машинного перевода: полный и частичный. При полном переводе переводится весь текст, а при частичном – только отдельные лексические единицы. При этом для перевода отбирается только такая лексика, которая не вошла в русскоязычные тональные словари. Результаты экспериментальных исследований по выявлению наиболее эффективного метода (показывающего наибольшую точность) к машинному переводу русскоязычных текстовых данных на английский язык для распознавания сентимента и эмоций представлены в таблице 4.

Из результатов экспериментов видно, что использование полного перевода на английский язык является более эффективным методом для распознавания сентимента и эмоций только в экспертных транскрипциях. Такой метод позволяет достичь точности классификации семи классов эмоций и трех классов сентимента по показателю взвешенной F-меры 31,12% и 75,37%, соответственно. Полученные результаты превосходят другие методы анализа текстовых транскрипций (использование оригинального текста с русскими тональными словарями и частичный перевод с английскими словарями) на 5-10% по показателю взвешенной F-меры.

Таблица 4. Результаты распознавания сентимента и эмоций при использовании различных методов машинного перевода, взвешенная F-мера (%)

Транскрипции	**Тип перевода**	**Эмоции**	**Сентимент**
Экспертные	Русскоязычный текст	27,42	65,61
	Полный перевод	**31,12**	**75,37**
	Частичный перевод	26,21	71,79
Автоматические	Русскоязычный текст	23,09	67,79
	Полный перевод	22,78	66,62
	Частичный перевод	**23,74**	**71,60**

Для автоматических транскрипций корпуса RAMAS наиболее эффективным методом достичь наибольшей точности распознавания сентимента и эмоций является использование частичного перевода на английский язык. Такие результаты могут быть обоснованы низким качеством автоматических транскрипций, которые в свою очередь зависимы от качества записи корпуса данных. Можно предположить, что наличие фраз сторонних дикторов в анализируемом высказывании сильно зашумляет эмоциональную окраску всего текста.

5.3. Репрезентативные признаки. В настоящем исследовании предлагается извлекать текстовые признаки при помощи тональных словарей. В предложенном методе вектор транскрипций состоит из статистического вектора и тонального вектора. Статистический вектор представляет собой набор статистических показателей всего анализируемого высказывания и состоит из следующих параметров: количество положительных, отрицательных, нейтральных слов, среднее значение сентимента всех слов, сумма весов положительных и отрицательных слов, максимальное и минимальное значение весов (всего восемь значений). Тональный вектор содержит информацию о последовательности эмоциональных слов в высказывании.

В экспериментах выше использовался конкатенированный вектор из статистического и тонального векторов. Для приведения всех векторов к единой длине все вектора либо обрезались, либо дополнялись нулями до значения 98 (8 – статистический вектор и 90 – тональный вектор). В данном экспериментальном исследовании исследуется гипотеза о том, является ли статистический вектор полностью репрезентативным для анализа сентимента. Под репрезентативностью понимается содержание характеристик в малой выборке, отражающей характеристики генеральной совокупности. Суть эксперимента заключалась в следующем: использовалось различное количество показателей статистического вектора (от двух до восьми), использовались различные комбинации статистических показателей, с каждой комбинацией статистических показателей векторов строился классификатор распознавания сентимента для всех типов данных (оригинальные русскоязычные тексты, полный и частичный переводы на английский язык), затем результаты усреднялись по комбинациям в зависимости от количества использованных статистических показателей. Полученные результаты представлены на рисунке 6 в виде гистограммы, по оси абсцисс которой располагаются количество параметров (в комбинации) в статистическом векторе для различных типов данных, по оси ординат – точность распознавания трех классов сентимента по показателю взвешенной F-меры.

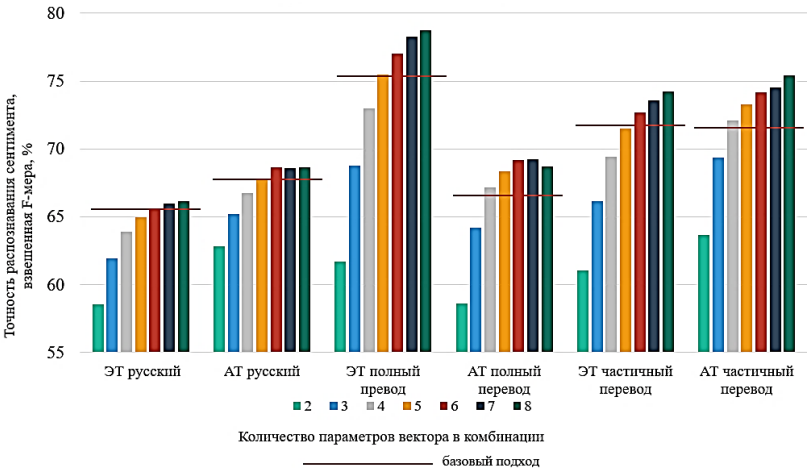

Рис. 6. Гистограмма влияния количества параметров статистического вектора на точность распознавания сентимента

На рисунке 6 красной линией отмечен базовый результат, полученный в ходе предыдущих экспериментальных исследований с использованием конкатенированного вектора (статистического и тонального) текстовых транскрипций, представленный в таблице 4. Как можно заметить из результатов экспериментов, использование только статистических показателей при анализе сентимента в текстовых данных позволяет достичь более высокой точности относительно использования конкатенированного вектора. В большинстве случаев, чем больше количество статистических показателей, тем выше точность распознавания сентимента. Однако, даже использование минимум шести статистических показателей позволяют превзойти базовый метод (в котором используется конкатенированный вектор) примерно на 1-5% по показателю взвешенной F-меры. Это может быть обосновано тем, что классификатор при анализе только статистических показателей сосредотачивается на более репрезентативной информации о сентименте. Под репрезентативностью понимается сконцентрированная информация о сентименте, находящаяся в статистическом векторе, когда тональный вектор содержит в себе разряженную информацию. Также можно сделать вывод о том, что последовательность эмоциональных слов (тональный вектор) для автоматического распознавания сентимента не важна. Более того, использование тонального вектора может только зашумлять информацию о тональности в высказываниях, что делает данный вектор

не репрезентативным для классификатора. Еще одним преимуществом использования только статистического вектора является сокращение времени обучения классификаторов распознавания сентимента. В предыдущих исследованиях авторов данной статьи [15] для распознавания сентимента использовался только тональный вектор на автоматических транскрипциях корпуса RAMAS, где достигнуто значение 43,31% взвешенной F-меры, что на 22% ниже использования только статистического вектора для аналогичной задачи.

Также в ходе экспериментов было установлено какие именно статистические показатели вносят больший вклад в распознавание сентимента в текстовых транскрипциях. Для выделения такой информации была использована оригинальная методика, описанная ниже. На рисунке 7 представлена гистограмма, по оси абсцисс которой представлены статистические показатели, а по оси ординат усредненная точность распознавания сентимента по показателю взвешенной F-меры в %. Усреднение точности происходило на основе предыдущего эксперимента по принципу усреднения полученной точности распознавания сентимента для каждого вектора, в котором встречался анализируемый показатель. Например, параметр количество негативных слов в предыдущем эксперименте встречался в различных M векторах. Для получения усредненной точности рассматриваемого параметра вычислялись точности распознавания сентимента с использованием каждого из векторов M и находилось среднее арифметическое этих значений точности.

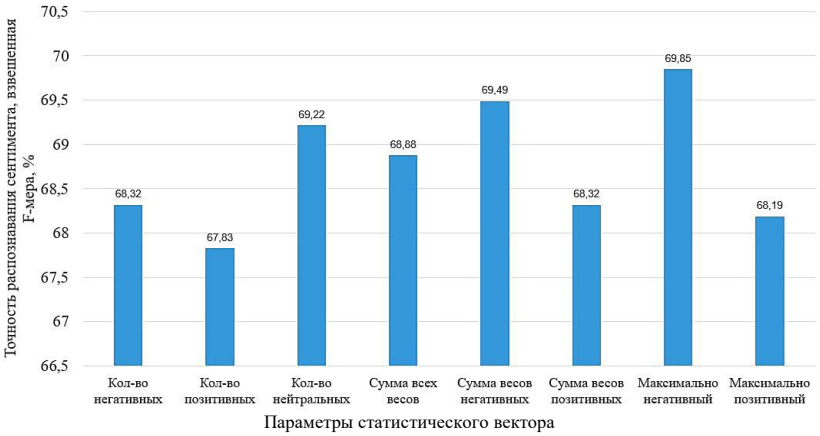

Рис. 7. Гистограмма вклада параметров статистического вектора в точность распознавания сентимента

Гистограмма на рисунке 7 свидетельствует о том, что каждый статистический показатель достаточно важен для распознавания сентимента. Так, разброс между крайними значениями показателей составляет менее 2% (в диапазоне от 67,75 до 69,75%). При этом такие показатели как максимальный вес негативного слова, сумма весов негативных слов и количество нейтральных слов вносят больший вклад в распознавание сентимента в текстовых транскрипциях речевых высказываний.

6. Обсуждение результатов. Вследствие проведенных экспериментов получены несколько теоретических и эмпирических результатов, которые имеют значение в контексте автоматического анализа текстов на естественных языках. Во-первых, лемматизация оказалась более предпочтительным методом, повышая точность распознавания как для экспертных транскрипций, так и для автоматических. По-видимому, лемматизация позволяет обеспечить лучший результат распознавания нормализованного текста за счет того, что учитывает контекст и частеречную принадлежность слова. Кроме того, лемматизация позволяет снизить количество ошибок, которые могут возникнуть из-за неправильного обрезания слов при стемминге. С другой стороны, лемматизация, в отличие от стемминга, требует больших временных и вычислительных ресурсов.

Другим важным результатом является возможность объединения тональных словарей. Тональные словари являются одним из определяющих факторов для определения сентимента текстов, однако могут быть неполными и/и не учитывать различные факторы. Объединение данных из различных источников предсказуемо повышает количество обучающих данных. В то же время, привлечение языкового материала на иностранных языках повышает качество распознавания по той причине, что основным способом выражения оценочной модальности (эмоциональной окраски текста) в языках мира является именно лексика, а не синтаксис или морфология. Лексические единицы, маркированные в отношении оценочной модальности, проявляют устойчивость, совпадая по тональности в различных языках. Таким образом, машинный перевод оказывается эффективным инструментом, позволяющим пополнять обучающие данные напрямую из других языков. Следует иметь в виду, что однозначный машинный перевод с одного естественного языка на другой пока невозможен, поэтому машинный перевод должен, в идеале, сопровождаться экспертной проверкой. Так или иначе, даже прямолинейное применение машинного перевода в представленных в рамках настоящей статьи экспериментов позволило поднять значение

F-меры распознавания на 1-4% для эмоций, и на 5-10% для сентимента.

Наконец, последним важным выводом является возможность применения статистического вектора на этапе векторизации текста вместо конкатенированного тонального и статистического. В самом деле, объединение тональных и статистических векторов может увеличить размерность векторного пространства, что может привести к проблемам с памятью и производительностью модели определения тональности. Как показали эксперименты, использование только статистических векторов может быть более эффективным для достижения большей точности распознавания сентимента, а также уменьшения времени автоматического анализа текстовых данных.

7. Заключение. Статья посвящена разработке метода распознавания сентимента и эмоций в орографических транскрипциях речи дикторов. Предложенный метод основан на использовании тональных словарей для векторизации текстов, а также машинного перевода русскоязычных транскрипций на английский язык. В статье описываются экспериментальные исследования с методами предобработки текстовых данных, машинного перевода (полного и частичного), выделения репрезентативных признаков из тональных векторов.

По результатам проведенного исследования можно сделать следующие выводы:

1) Лемматизация в качестве метода нормализации слов в текстовых транскрипциях речи позволяет повысить точность распознавания сентимента и эмоций.

2) Полный перевод на английский язык русскоязычных экспертных транскрипций речи увеличивает точность распознавания эмоций (семь классов) и сентимента (три класса).

3) Для автоматических транскрипций речи частичный перевод на английский язык помогает увеличить точность распознавания сентимента и эмоций.

4) Использование статистического вектора в качестве векторизации текстовых транскрипций речи позволяет получить точность выше относительно использование конкатенированного (статистического и тонального) вектора.

Для достижения наилучших результатов в распознавании эмоций в русскоязычных текстах необходимо комбинировать различные методы и подходы, включая машинный перевод и использование объединенных тональных словарей. Важно также учитывать специфику текста и задачу, которую необходимо решить,

при выборе подходящего метода. В целом, исследования в области анализа сентимента и эмоций в русскоязычных текстах необходимо продолжать с целью поиска эффективных методов, с помощью которых можно достичь наибольшей точности.

Литература

1. Николаев И.С., Митренина О.В., Ландо Т.М. Прикладная и компьютерная лингвистика // М.:ЛЕНАНД. 2017. 320 с.
2. Carosia A.E.O., Coelho G.P., Silva A.E.A. Analyzing the Brazilian financial market through portuguese sentiment analysis in social media // Applied Artificial Intelligence. 2020. vol. 34. no. 1. pp. 1–19.
3. Smetanin S. The applications of sentiment analysis for Russian language texts: Current challenges and future perspectives // IEEE Access. 2020. vol. 8. pp. 110693–110719. DOI: 10.1109/ACCESS.2020.3002215.
4. Карпов А.А., Юсупов Р.М. Многомодальные интерфейсы человеко-машинного взаимодействия // Вестник Российской академии наук. 2018. Т. 88. № 2. С. 146–155.
5. Dvoynikova A., Verkholyak O., Karpov A. Analytical review of methods for identifying emotions in text data // CEUR-WS. 2020. vol. 2552. pp. 8–21.
6. Ekman P. An Argument for Basic Emotions // Cognition and Emotion. 1992. vol. 6(3-4). pp. 169–200.
7. Dvoynikova A., Karpov A. Bimodal sentiment and emotion classification with multi-head attention fusion of acoustic and linguistic information // Computational Linguistics and Intellectual Technologies. 2023. vol. 22. pp. 51–61.
8. Viksna R., Jekabsons G. Sentiment analysis in Latvian and Russian: A survey // Applied Computer Systems. 2018. vol. 23. no. 1. pp. 45–51.
9. Loukachevitch N., Levchik A. Creating a general Russian sentiment lexicon // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. pp. 1171–1176.
10. Demirtas E., Pechenizkiy M. Cross-lingual polarity detection with machine translation // Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. 2013. pp. 1–8.
11. Reichel J., Benko Ľ. The Influence of a Machine Translation System on Sentiment Levels // RASLAN 2022 Recent Advances in Slavonic Natural Language Processing. 2022. pp. 201–208.
12. Zygadlo A., Kozlowski M., Janicki A. Text-Based emotion recognition in English and Polish for therapeutic chatbot // Applied Sciences. 2021. vol. 11(21). no. 10146.
13. Nandwani P., Verma R. A review on sentiment analysis and emotion detection from text // Social Network Analysis and Mining. 2021. vol. 11(1). no. 81.
14. Hartung K., Herygers A., Kurlekar S.V., Zakaria K., Volkan T., Gröttrup S., Georges M. Measuring Sentiment Bias in Machine Translation // International Conference on Text, Speech, and Dialogue. 2023. pp. 82–93.
15. Двойникова А.А. Сентимент-анализ транскрипции разговорной речи при помощи автоматического машинного перевода // Сборник трудов IX Конгресса молодых ученых. 2021. С. 199–203.
16. Perepelkina O., Kazimirova E., Konstantinova M. RAMAS: Russian Multimodal Corpus of Dyadic Interaction for studying emotion recognition // PeerJ Preprints. 2018. vol. 6. no. e26688v1.
17. Russell J.A. A circumplex model of affect // Journal of personality and social psychology. 1980. vol. 39. no. 6. pp. 1161–1178.

18. Плунгян В.А. Введение в грамматическую семантику: Грамматические значения и грамматические системы языков мира // М.: РГГУ. 2011. 672 с.
19. Goddard C., Wierzbicka A. Semantic and Lexical Universals // Studies in Second Language Acquisition, 1996. vol. 18(4). 520 p.
20. Котельников Е.В., Разова Е.В., Котельникова А.В., Вычегжанин С.В. Современные словари оценочной лексики для анализа мнений на русском и английском языках (аналитический обзор) // Научно-техническая информация. Серия. 2020. Т. 2. С. 16–33.
21. Hercig T., Brychcín T., Svoboda L., Konkol M. Uwb at semeval-2016 task 5: Aspect based sentiment analysis // Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). 2016. pp. 342–349.
22. Koltsova O.Y., Alexeeva S., Kolcov S. An opinion word lexicon and a training dataset for Russian sentiment analysis of social media // Computational Linguistics and Intellectual Technologies. 2016. vol. 15. pp. 277–287.
23. Strapparava C., Valitutti A. Wordnet affect: an affective extension of wordnet // Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC). 2004. pp. 1083–1086.
24. Mohammad S.M., Turney D.P. Crowdsourcing a word-emotion association lexicon // Computational Intelligence. 2013. vol. 29(3). pp. 436–465.
25. Hu M., Liu B. Mining and summarizing customer reviews // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004. pp. 168–177.
26. Hutto C., Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text // Proceedings of the international AAAI conference on web and social media. 2014. vol. 8. no. 1. pp. 216–225.

Двойникова Анастасия Александровна — младший научный сотрудник, лаборатория речевых и многомодальных интерфейсов, Федеральное государственное бюджетное учреждение науки "Санкт-Петербургский Федеральный исследовательский центр Российской академии наук". Область научных интересов: искусственный интеллект, машинное обучение, нейронные сети, сентимент-анализ, анализ аффективных состояний человека. Число научных публикаций — 20. dvoynikova.a@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Кагиров Ильдар Амирович — научный сотрудник, лаборатория речевых и многомодальных интерфейсов, Федеральное государственное бюджетное учреждение науки "Санкт-Петербургский Федеральный исследовательский центр Российской академии наук". Область научных интересов: корпусная лингвистика, малоресурсные языки. Число научных публикаций — 40. kagirov@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Карпов Алексей Анатольевич — д-р техн. наук, профессор, руководитель лаборатории, лаборатория речевых и многомодальных интерфейсов, Федеральное государственное бюджетное учреждение науки "Санкт-Петербургский Федеральный исследовательский центр Российской академии наук". Область научных интересов: речевые технологии, автоматическое распознавание речи, обработка аудиовизуальной речи, многомодальные человеко-машинные интерфейсы, компьютерная паралингвистика и другие. Число научных публикаций — 350. karpov@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Поддержка исследований. Раздел 4 выполнен в рамках бюджетной темы СПб ФИЦ РАН (№ FFZF-2022-0005), остальные исследования выполнены при финансовой поддержке Российского научного фонда, проект № 22-11-00321.

DOI 10.15622/ia.23.4.9

A. DVOYNIKOVA, I. KAGIROV, A. KARPOV

A METHOD FOR RECOGNITION OF SENTIMENT AND EMOTIONS IN RUSSIAN SPEECH TRANSCRIPTS USING MACHINE TRANSLATION

Dvoynikova A., Kagirov I., Karpov A. **A Method for Recognition of Sentiment and Emotions in Russian Speech Transcripts Using Machine Translation.**

Abstract. This paper addresses the issue of user emotions and sentiment recognition in transcripts of Russian speech samples using lexical methods and machine translation. The availability of data for sentiment analysis in Russian texts is quite limited, thus this paper proposes a new approach which is based on automatic machine translation of Russian texts into English. Additionally, the paper presents the results of experimental research regarding the impact of partial and full machine translation on emotion and sentiment recognition. Partial translation means translating single lexemes not included in Russian sentiment dictionaries, while full translation implies translating the entire text. A translated text is further analyzed using different English sentiment dictionaries. Experiments have demonstrated that the combination of all English sentiment dictionaries enhances the accuracy of emotion and sentiment recognition in text data. Furthermore, this paper explores the correlation between the length of the text data vector and its representativity. Experimental research for emotion and sentiment recognition tasks was conducted with the use of expert and automatic transcripts of the multimodal Russian corpus RAMAS. Based on the experimental results, one can conclude that the use of word lemmatization is a more effective approach for normalizing words in speech transcripts compared to stemming. The use of the proposed methods involving full and partial machine translation allows for an improvement in sentiment and emotion recognition accuracy by 0.65-9.76% in terms of F-score compared to the baseline approach. As a result of the application of machine translation methods to expert and automatic transcriptions of the Russian speech corpus RAMAS, an accuracy in recognition of 7 emotion classes was achieved at 31.12% and 23.74%, and 3 sentiment classes at 75.37% and 71.60%, respectively. Additionally, the experiments revealed that the use of statistical vectors as a text data vectorization method results in an a 1-5% increase in F-score value compared to concatenated (statistical and sentiment) vectors.

Keywords: machine translation, sentiment dictionaries, emotion recognition, sentiment analysis, sentiment vectors.

References

1. Nikolaev I.S., Mitrenina O.V., Lando T.M. Prikladnaya i komp'yuternaya lingvistika [Applied and computational linguistics]. M.: LENAND, 2017. 320 p (In Russ.).
2. Carosia A.E.O., Coelho G.P., Silva A.E.A. Analyzing the Brazilian financial market through portuguese sentiment analysis in social media. Applied Artificial Intelligence. 2020. vol. 34. no. 1. pp. 1–19.
3. Smetanin S. The applications of sentiment analysis for Russian language texts: Current challenges and future perspectives. IEEE Access. 2020. vol. 8. pp. 110693–110719. DOI: 10.1109/ACCESS.2020.3002215.
4. Karpov A.A., Yusupov R.M. Multimodal interfaces of human-computer interaction, Herald of the Russian Academy of Sciences. 2018. vol. 88. no. 2. pp. 146–155.
5. Dvoynikova A., Verkholyak O., Karpov A. Analytical review of methods for identifying emotions in text data. CEUR-WS. 2020. vol. 2552. pp. 8–21.

6. Ekman P. An Argument for Basic Emotions. Cognition and Emotion. 1992. vol. 6(3-4). pp. 169–200.
7. Dvoynikova A., Karpov A. Bimodal sentiment and emotion classification with multi-head attention fusion of acoustic and linguistic information. Computational Linguistics and Intellectual Technologies. 2023. vol. 22. pp. 51–61.
8. Viksna R., Jekabsons G. Sentiment analysis in Latvian and Russian: A survey. Applied Computer Systems. 2018. vol. 23. no. 1. pp. 45–51.
9. Loukachevitch N., Levchik A. Creating a general Russian sentiment lexicon. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. pp. 1171–1176.
10. Demirtas E., Pechenizkiy M. Cross-lingual polarity detection with machine translation. Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. 2013. pp. 1–8.
11. Reichel J., Benko Ľ. The Influence of a Machine Translation System on Sentiment Levels. RASLAN 2022 Recent Advances in Slavonic Natural Language Processing. 2022. pp. 201–208.
12. Zygadlo A., Kozlowski M., Janicki A. Text-Based emotion recognition in English and Polish for therapeutic chatbot. Applied Sciences. 2021. vol. 11(21). no. 10146.
13. Nandwani P., Verma R. A review on sentiment analysis and emotion detection from text // Social Network Analysis and Mining. 2021. vol. 11(1). no. 81.
14. Hartung K., Herygers A., Kurlekar S.V., Zakaria K., Volkan T., Gröttrup S., Georges M. Measuring Sentiment Bias in Machine Translation. International Conference on Text, Speech, and Dialogue. 2023. pp. 82–93.
15. Dvoynikova A.A. [Sentiment analysis of transcription of spoken speech using automatic machine translation] Sbornik trudov IX Kongressa molodyh uchenyh [Proceedings of the IX Congress of Young Scientists]. 2021. pp. 199–203. (In Russ.).
16. Perepelkina O., Kazimirova E., Konstantinova M. RAMAS: Russian Multimodal Corpus of Dyadic Interaction for studying emotion recognition. PeerJ Preprints. 2018. vol. 6. no. e26688v1.
17. Russell J.A. A circumplex model of affect. Journal of personality and social psychology. 1980. vol. 39. no. 6. pp. 1161–1178.
18. Plungjan V.A. Vvedenie v grammaticheskuju semantiku: Grammaticheskie znachenija i grammaticheskie sistemy jazykov mira [Introduction to Grammatical Semantics: Grammatical Meanings and Grammatical systems of the languages of the world]. M.: RSUH. 2011. 672 p (In Russ.).
19. Goddard C., Wierzbicka A. Semantic and Lexical Universals. Studies in Second Language Acquisition, 1996. vol. 18(4). 520 p.
20. Kotel'nikov E.V., Razova E.V., Kotel'nikova A.V., Vychegzhanin S.V. [Modern dictionaries of evaluation vocabulary for the analysis of opinions in Russian and English (analytical review)]. Nauchno-tehnicheskaja informacija. Serija – Scientific and technical information. Series. 2020. vol. 2. pp. 16–33 (In Russ.).
21. Hercig T., Brychcín T., Svoboda L., Konkol M. Uwb at semeval-2016 task 5: Aspect based sentiment analysis. Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). 2016. pp. 342–349.
22. Koltsova O.Y., Alexeeva S., Kolcov S. An opinion word lexicon and a training dataset for Russian sentiment analysis of social media. Computational Linguistics and Intellectual Technologies. 2016. vol. 15. pp. 277–287.
23. Strapparava C., Valitutti A. Wordnet affect: an affective extension of wordnet. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC). 2004. pp. 1083–1086.
24. Mohammad S.M., Turney D.P. Crowdsourcing a word-emotion association lexicon. Computational Intelligence. 2013. vol. 29(3). pp. 436–465.

25. Hu M., Liu B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004. pp. 168–177.
26. Hutto C., Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the international AAAI conference on web and social media. 2014. vol. 8. no. 1. pp. 216–225.

Dvoynikova Anastasia — Junior researcher, Laboratory of speech and multimodal interfaces, St. Petersburg Federal Research Center of the Russian Academy of Sciences. Research interests: artificial intelligence, machine learning, neural networks, sentiment analysis, human affective states analysis. The number of publications — 20. dvoynikova.a@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Kagirov Ildar — Researcher, Laboratory of speech and multimodal interfaces, St. Petersburg Federal Research Center of the Russian Academy of Sciences. Research interests: corpus linguistics, low-resource languages. The number of publications — 40. kagirov@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Karpov Alexey — Ph.D., Dr.Sci., Professor, Head of the laboratory, Laboratory of speech and multimodal interfaces, St. Petersburg Federal Research Center of the Russian Academy of Sciences. Research interests: speech technology, automatic speech recognition, audio-visual speech processing, multimodal human-computer interfaces, and computational paralinguistics. The number of publications — 350. karpov@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Acknowledgements. Research presented in Section 4 was financially supported by a state research grant for SPC RAS (Topic No. FFZF-2022-0005), the remainder of the study was funded by RSF (Project 22-11-00321).