DOI 10.58351/2949-2041.2026.35.6.022

Бойченко Елена Викторовна, Магистрант
АНО ВО «Российский Новый Университет»
Boichenko Elena Viktorovna
Russian New University

Золотарев Олег Васильевич
кандидат технических наук, доцент
АНО ВО «Российский Новый Университет»
Zolotarev Oleg Vasilevich, PhD, Associate Professor
Russian New University

МЕТОДЫ АВТОМАТИЧЕСКОГО АНАЛИЗА ПСИХОЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ ПО ТЕКСТУ НА РУССКОМ И АНГЛИЙСКОМ ЯЗЫКАХ
METHODS OF AUTOMATIC PSYCHO-EMOTIONAL STATE ANALYSIS FROM RUSSIAN AND ENGLISH TEXTS

Аннотация. В статье рассматриваются методы автоматизированного определения психоэмоционального состояния человека по тексту на русском и английском языках. Обосновывается необходимость использования трансформер-моделей (RuBERT, mBERT) с учетом ограничений по ресурсам и наличию разметанных корпусов.

Abstract. The paper discusses approaches to the automatic detection of a person's psycho-emotional state from text in Russian and English. The paper substantiates the feasibility of using transformer models (RuBERT, mBERT), considering resource constraints and the availability of annotated corpora.

Ключевые слова: Психоэмоциональное состояние, анализ текста, обработка естественного языка, машинное обучение, глубокое обучение, BERT.

Keywords: Psycho-emotional state, text analysis, natural language processing, machine learning, deep learning, BERT.

Введение

Увеличение объема текстовой информации в социальных сетях, мессенджерах и онлайн-сервисах вызвало повышение интереса к автоматическим методам анализа эмоционального состояния пользователей по их текстовым сообщениям. Психоэмоциональное состояние проявляется в лексике, синтаксисе и жанровых особенностях текста, и оно может быть оценено средствами обработки естественного языка (NLP) и машинного обучения [4].

Одновременный анализ текстов на двух языках, в частности на русском и английском языках, представляет собой задачу, вызывающую особый интерес. Эти языки очень различаются по морфологии, синтаксису и даже по выражению чувств [7]. В работах по построению эмоциональных лексиконов и корпусов для английского и русского языков отмечается различие в структуре эмоциональной лексики и частотности использования [6].

С практической точки зрения анализ психоэмоционального состояния применяется в системах психологического консультирования, мониторинга токсичного контента, персонализации рекомендаций, а также в чат-ботах и голосовых ассистентах. С научной точки зрения актуальность темы обусловлена необходимостью сравнения эффективности различных методов для различных языков и типов текстов [1].

Цель статьи: теоретически обосновать выбор метода автоматизированного анализа психоэмоционального состояния по тексту на русском и английском языках.

Для этого необходимо:

– изучить существующие методы определения эмоций и психоэмоционального состояния по тексту;

– выявить особенности применения этих методов к русскому и английскому языкам;

– на основе открытых источников выполнить гипотетическое сравнение лексико-статистических методов, классических методов машинного обучения и методов глубокого обучения;

– обосновать выбор наиболее подходящего для реализации в рамках ресурсных и дата-доступностных ограничений метода.

Методом является изучение и обобщение научных публикаций в открытом доступе, сравнительный анализ известных моделей и гипотетическое моделирование их применения.

2. Теоретические основы анализа психоэмоционального состояния по тексту

2.1. Понятие психоэмоционального состояния и его лингвистические маркеры

Психоэмоциональное состояние представляет собой уровень развития психоэмоциональных реакций и ответов человека на воздействия экстремальной среды, с последующим влиянием на его речевую активность [4]. В письменном общении такие состояния выражаются в лексическом выборе, грамматической структуре, в использовании стилистических средств и методов организации текста.

Психоэмоциональное состояние [4] представляется в форме трёх типов лингвистических маркеров:

– лексический (эмоционально окрашенные слова и фразы);

– синтаксический (особенности конструкции предложения, применения вопросов, восклицаний, эллипс);

– семантический (сюжетные смыслы, символы, ирония, сарказм).

Русскому языку свойственно богатство фразеологизмов и устойчивых выражений для передачи эмоций («сердце ушло в пятки», «руки опускаются»), а также развитая система словообразования для формирования слов с эмоциональной окраской. Английский язык тоже предлагает огромные возможности для выражения эмоций, но значительная их часть воплощена в слэнг, сокращения, эмодзи, пунктуацию, что, в первую очередь характерно интернет-общению [6].

Эти отличия означают, что методы анализа, созданные для английского языка, не всегда могут быть напрямую применены к русскому, и им необходимо адаптироваться или обучаться на специализированных корпусах. При разработке мультиязычного подхода следует учитывать как универсальные принципы эмоционального анализа, так и специфические особенности анализируемых языков [5].

2.2. Лексико-статистические методы анализа

Диагностические лексико-статистические методы оценки психоэмоционального состояния базируются на предопределенных словарях, в которых словам/фразам приписана информация о валентности (позитивная/негативная), интенсивности и типе эмоции. Для анализа англоязычных текстов получил широкое распространение инструмент VADER (Valence Aware Dictionary and sEntiment Reasoner), который демонстрирует высокую результативность при анализе текстов из соцсетей с учетом особенностей интернет-сленга, заглавных букв и пунктуации [1].

Для русского языка разрабатываются собственные лексические ресурсы. В частности, в работе [6] приводятся данные об аннотированных корпусах русскоязычных текстов, а работа [2] демонстрирует для русского языка (RusEmoLex), представляющий информацию о валентной структуре и типе эмоции для нескольких лексических единиц. В ряде исследований также создаются специализированные словари для отдельных доменов анализа отзывов, новостей, комментариев [4].

С простотой реализации и малыми вычислительными затратами связаны и основные достоинства лексико-статистических методов. Их легко вставить в существующие программные системы и использовать в ограниченных условиях. Вместе с тем их точность и качество остаются на невысоком уровне, т.к. в неявном виде они достаточно слабо рассматривают контекст, полисемию, иронию и сарказм. По рейтингам существующих исследований, точность с классификаторами, основанных на словарных методах, доходил до 65–75% в зависимости от корпуса и языка [3].

2.3. Методы машинного обучения

Машино обучающиеся алгоритмы позволяют перейти от статичных словарей к моделям, основанным на размеченных данных, тем самым учитывая более сложные взаимосвязи между признаками текста и целевыми эмоциональными метками. Наивный байесовский классификатор, метод опорных векторов (SVM) и логистическая регрессия – это некоторые из моделей, которые традиционно используются в задачах анализа тональности и эмоций [5].

В качестве признаков, как правило, используются: мешок слов (bag-of-words), n-граммы, TF-IDF-веса, лексические и синтаксические характеристики текста и другие статистические подходы.

Для английского присутствуют разнообразные открытые наборы с разметками эмоций и тональности с обзора, основанные на рецензиях, твитах и заголовках новостей [7]. Для русского языка обзор доступных наборов данных представлен в обзоре [4], в том числе включен сравнительный анализ различных алгоритмов. Было показано, что при достаточной настройке и выборе признаков наивные модели способны к точности 80–85% для отдельных задач.

Преимуществом таких моделей является то, что их можно довольно быстро и интерпретируемо адаптировать к конкретному предметному полю, даже при небольшом количестве данных. Их минус в том, что они плохо учились на сложные контекстуальные зависимости, а переносимость моделей между языками и жанрами была невысокой.

2.4. Модели глубокого обучения и трансформеры

Развитие методов глубокого обучения позволило создавать модели, учитывающие контекст всего текста и взаимосвязи между словами в предложении. Ключевую роль в этом сыграла архитектура Transformer, на основе которой построены модели семейства BERT и их многоязычные модификации. Для английского языка базовой моделью является BERT, предварительно обученная на больших текстовых корпусах и затем дообучаемая на прикладных задачах, включая анализ тональности. Для многоязычных сценариев применяется mBERT, а для русского языка – RuBERT. Исследования показывают, что RuBERT эффективно применяется в задачах sentiment analysis и emotion detection, превосходя классические лексико-статистические подходы по качеству классификации.

Дополнительное преимущество обеспечивают мультиязычные модели – mBERT и XLM-R, позволяющие обрабатывать тексты на нескольких языках в рамках единой архитектуры. Это особенно актуально для задач, в которых необходим консистентный анализ субъективной оценки эмоциональной окраски как русскоязычных, так и англоязычных текстов. Недостатками моделей глубокого обучения являются большая потребность вычислительных ресурсов, необходимость использования графических процессоров (GPU) для эффективного обучения и применения, а также сложность реализации и отладки. Тем не менее, при высоких требованиях к точности, и наличии современной программного-аппаратной базы, модели на основе трансформеров признаются наиболее перспективным подходом.

3. Особенности анализа психоэмоционального состояния на русском и английском языках

3.1. Лингвистические различия и их влияние на анализ

Русский относится к флективным языкам с развитой системой склонения и спряжения, благодаря которой образуется огромное количество словоформ, требующих нормализации (лемматизации) до анализа [9]. Для некоторых задач это затрудняет создание словарей и статистических моделей: без нормализации признаки являются избыточными и разреженными. С другой стороны, английский язык является более аналитическим, имеет менее богатую морфологию и в целом характеризуется более устойчивыми формами, что облегчает разработку моделей на уровне слов.

Способы выражения эмоций в русском и английском заметно отличаются. В исследованиях лексики эмоций выявляются различия в частоте положительных и отрицательных слов, а также в степени выраженности тех или иных эмоций [6]. Это подразумевает, что прямое перенесение словарей и обученных систем с одного языка на другой без адаптации к лексическим и культурным особенностям может привести к снижению точности.

3.2. Корпусные ресурсы и эмоциональные лексиконы

Условием эффективного применения методов анализа психоэмоционального состояния является наличие качественных корпусов и лексиконов эмоций, что особенно важно при работе с многоязычными или культурно специфичными данными. Для русского языка одним из актуальных проектов является RusEmoLex – лексикон эмоций, построенный на основе корпусных данных и экспертной разметки [2]. Систематизация существующих русскоязычных датасетов с описанием их объемов, типов разметки и жанров текстов представлена в исследовании [4].

Для английского языка существует огромное количество открытых корпусов, в том числе параллельные корпуса отзывов и рецензий, примененные, в частности, при построении сопоставимых русско-английских корпусов по анализу субъективности [5]. Эти наборы ресурсов в совокупности дают возможность исследователям для сравнения методов двух языков, а также для оценки переносимости моделей.

3.3. Сопоставление эмоциональных концептов

Русский и английский в сравнительном аспекте эмоциональных концептов на сегодня исследованы достаточно, в том числе и исходя из положения, что базовые эмоции совпадают, однако их лингвистическое воплощение и частотность употребления могут существенно различаться [7]. Это создает дополнительное ограничение для моделей, обученных на одном лишь языке, поскольку их модели восприятия оказываются культура- и язык-специфичными.

Мультиязычные трансформер-модели отчасти решают эту проблему, так как обучаются на многоязычных корпусах, и имеют возможность формировать абстрактные (общие) представления в разных языках. Тем не менее, для работы, где требуется достаточно точная интерпретация и детализация эмоций, особенно в психолингвистическом аспекте, может понадобиться дополнительная адаптация и локальное дообучение моделей [8].

4. Методология гипотетического сравнительного исследования

В статье описывается гипотетическое исследование, эффективность которого заключается в сравнении трёх классов методов: лексико-статистических, традиционных методов машинного обучения и моделей глубокого обучения на базе трансформеров.

В гипотетическом сценарии исследователь имеет корпус, состоящий из пользовательских сообщений и отзывов (социальные сети и тематические форумы), размеченных по эмоциональным классам, на русском и английском языках. Корпуса, подобные тем, что описаны в [4] и [5] (описания которых в предыдущем разделе).

Для оценки технологий по уровню соответствия были использованы следующие критерии: точность и полнота (precision, recall, F1-measure) для классификации эмоционального состояния; устойчивость к шуму и сленгу; поддержка многоязычного анализа; вычислительная сложность и требования ресурсов; легкость интеграции в программное обеспечение. Поскольку исследование гипотетическое, количественные значения заимствованы из опубликованных работ [1–5, 7] и обобщены для цели сравнения методов в одном формате.

5. Гипотетическое сравнительное исследование методов

5.1. Сравнение лексико-статистических методов

Предположим, что для русского языка используется лексикон RusEmoLex [2], а для английского – структурно похожий эмоциональный словарь с информацией о валентности и интенсивности. Тексты проходят предварительную токенизацию и нормализацию (для русского языка – лемматизация), затем вычисляют агрегированную эмоциональную оценку по найденным маркерам.

Согласно результатам, описанным в исследованиях по составлению эмоциональных лексиконов [3], ожидается, что точность многоклассовой классификации (положительная, отрицательная, нейтральная, несколько базовых эмоций) будет на уровне 65-75%, в зависимости от домена и качества текстов. Наряду с тем наблюдается понижение точности на текстах, содержащих иронию, сарказм, смешанные эмоции.

5.2. Сравнение методов машинного обучения

Для машинного обучения рассматриваются алгоритмы SVM и логистической регрессии с признаковыми описаниями, получаемыми с помощью TF-IDF и n-грамм. Обучение проводится отдельно для русского и английского корпусов. На базе интеграции опубликованных результатов [4, 7] ожидается, что подобного рода модели будут иметь точность порядка 80–85% в особенности для задач бинарной классификации тональности. Преимуществом такого подхода является его относительная простота в обучении, а также возможность адаптации к различным жанрам и доменам при наличии соответствующих размеченных данных. Вместе с тем инвариантность и мультилингвистическая обработка смешанных текстов по-прежнему ограниченные: модели, обученные на одном языке, должны быть переобучены для другого.

5.3. Сравнение моделей глубокого обучения

RuBERT для русского и мультиязычная модель mBERT либо XLM-R для русских и английских текстов выступают в роли моделей глубокого обучения. Таким образом эти модели считаются предварительно обученными на больших корпусах и дополнительное обучение на задачах классификации эмоций проводится с использованием соответсвующих размеченных данных.

Согласно опубликованным данным [1, 3], такие модели демонстрируют F1 меру 85-90% и выше для задач определения эмоций и анализа настроений, значительно превосходя традиционные методы и лексико-статистические подходы. Особенно ярко проявляется превосходство трансформер-моделей на текстах с сильными контекстуальными зависимостями и в случаях, когда необходимо учитывать многозначность лексики.

5.4. Обобщающие выводы гипотетического сравнения

Итоговые результаты гипотетического сравнения будут выглядеть так:

– лексико-статистические методы: алгоритмы просты в реализации и не требуют большого количества ресурсов, но обладают довольно ограниченной точностью и слабой учитываемостью контекста;

– традиционные методы машинного обучения: более точные, чем словарные, требуют размеченных корпусов, но ограничены возможностью учитывать контекст и в переносимости между языками;

– модели глубокого обучения на базе трансформеров: показывают наилучшие результаты по точности и устойчивости к различным видам текста, но требуют больших вычислительных ресурсов и более сложной реализации.

6. Обоснование выбора метода для проведения исследований

Разрабатываемое приложение должно получить в качестве входных данных текстовые сообщения на русском и английском языках и выдать оценку психоэмоционального состояния пользователя. При этом необходимо достичь разумного баланса между точностью анализа, необходимыми ресурсами и уровнем сложности реализации.

На основе выполненного гипотетического анализа и результатов открытых исследований [1–5, 7] наиболее перспективным является применение трансформер-моделей на основе глубокого обучения – RuBERT для русского языка, и многоязычных моделей для использования русских и английских текстов – mBERT. При ограниченной вычислительной мощности дистиллированные версии трансформер-моделей могут быть использованы либо комбинированный подход – лексические методы для первичного грубого отброса, а трансформер-модель – для уточнения.

Реализованные модели также могут быть внедрены в полноценное приложение как модуль текстовой обработки. Такая архитектура позволяет отделить обработку анализа от остального приложения и упростить замену или ре-факторинг компонента. Таким образом, выбор в пользу трансформерных моделей, учитывая их мультиязычность и качество анализа, является оправданным для реализации в дипломной работе при наличии грамотного распределения вычислительных ресурсов и использования открытых инструментов и библиотек.

Заключение

В статье приведён обзор основных методов по автоматическому анализу психоэмоционального состояния на русском и английском языках по тексту. Демонстрируется, что методы лексико-статистического анализа [3] имеют преимущества в сравнительной простоте реализации и доступности, однако они ограничены в точности, поскольку не могут полностью учитывать контекст и сложные языковые конструкции. Машинное обучение традиционными алгоритмами [4, 7] также даёт возможность улучшить качество разбора, используя статистические признаки и размеченный корпус, но и эти алгоритмы не лишены ограничений при обработке многозначной и сильно контекстуальной информации.

Современные глубокие модели на базе архитектуры трансформеров [1, 3] показывают лучшую точность и более высокую стабильность в различных наборах текстов, позволяют решать задачи мультилингвальной обработки, что является крайне важным при анализе русских и английских текстов в параллель. Проведенное гипотетическое сравнительное исследование позволяет обоснованно выбрать данные модели в качестве фундаментальных для разработки программного приложения.

Впоследствии намечается переход от моделирования к практике: выбор отдельной модели либо комплекса моделей, подготовка и разметка корпуса, проектирование и тестирование программы, а также измерение качества полученных результатов при помощи полевых данных. Результаты исследования могут быть использованы при создании систем мониторинга эмоционального состояния пользователей, поддержки психологического консультирования и обработки обратной связи в прикладных областях

Список литературы:

1. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186. – URL: https://arxiv.org/abs/1810.04805.

2. Golubev, A. Improving Results on Russian Sentiment Datasets / A. Golubev, N. Loukachevitch // Proceedings of RANLP. – 2020. – URL: https://arxiv.org/pdf/2007.14310.

3. Hutto, C. J. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text / C. J. Hutto, E. Gilbert // Proceedings of ICWSM. – 2014. – Vol. 8, No. 1. – P. 216–225. – URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

4. Kotelnikov, E. V. Current Landscape of the Russian Sentiment Corpora / E. V. Kotelnikov // arXiv. – 2021. – URL: https://arxiv.org/abs/2106.14434.

5. Smetanin, S. Deep Transfer Learning Baselines for Sentiment Analysis in Russian / S. Smetanin, M. Komarov // Information Processing & Management. – 2021. – Vol. 58, No. 3. – URL: https://arxiv.org/abs/2104.12986.

6. Smetanin, S. RuSentiTweet: A Sentiment Analysis Dataset of General Domain Tweets in Russian / S. Smetanin // PeerJ Computer Science. – 2022. – URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9454938/.

7. Zagibalov, T. Comparable English-Russian Book Review Corpora for Sentiment Analysis / T. Zagibalov, E. Belyatskaya, J. Carroll // Proceedings of WASSA Workshop, NAACL HLT. – 2010. – P. 128–136. – URL: https://users.sussex.ac.uk/~johnca/papers/wassa10.pdf.

8. Conneau, A. Unsupervised Cross-lingual Representation Learning at Scale / A. Conneau, K. Khandelwal, N. Goyal [et al.] // Proceedings of ACL. – 2020. – P. 8440–8451. – URL: https://arxiv.org/abs/1911.02116