

Neural Network-Based Sentiment Classification of Russian Sentences into Four Classes

M. A. Kosterin¹, I. V. Paramonov¹

DOI: [10.18255/1818-1015-2022-2-116-133](https://doi.org/10.18255/1818-1015-2022-2-116-133)

¹P. G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.

MSC2020: 68T50

Research article

Full text in Russian

Received April 28, 2022

After revision May 23, 2022

Accepted May 25, 2022

The paper is devoted to the classification of Russian sentences into four classes: positive, negative, mixed, and neutral. Unlike the majority of modern study in this area, the mixed sentiment class is introduced. Mixed sentiment sentences contain positive and negative sentiments simultaneously.

To solve the problem, the following tools were applied: the attention-based LSTM neural network, the dual attention-based GRU neural network, the BERT neural network with several modifications of the output layer to provide classification into four classes. The experimental comparison of the efficiency of various neural networks were performed on three corpora of Russian sentences. Two of them consist of users' reviews: one with wear reviews and another with hotel reviews. The third corpus contains news from Russian media. The highest weighted F-measure in experiments (0.90) was achieved when using BERT on the wear reviews corpus, as well as the highest weighted F-measure for positive and negative sentences (0.92 and 0.93, respectively). The best classification results for neutral and mixed sentences were achieved on the news corpus. For them F-measure was 0.72 and 0.58, respectively. As a result of experiments, the significant superiority of the BERT transfer network was demonstrated in comparison with older neural networks LTSM and GRU, especially for classification of sentences with weakly expressed sentiments. The error analysis showed that "adjacent" (positive/negative and mixed) classes are worse classified with BERT than "opposite" classes (positive and negative, neutral and mixed).

Keywords: sentiment analysis; neural network-based classifier; BERT; natural language processing

INFORMATION ABOUT THE AUTHORS

Maksim Alekseevich Kosterin | orcid.org/0000-0002-8298-3156. E-mail: makcost@gmail.com
post-graduate student.

Ilya Vyacheslavovich Paramonov | orcid.org/0000-0003-3984-8423. E-mail: ilya.paramonov@fruct.org
correspondence author | PhD, associate professor.

Funding: This work was supported by P. G. Demidov Yaroslavl State University Project No. VIP-016.

For citation: M. A. Kosterin and I. V. Paramonov, "Neural Network-Based Sentiment Classification of Russian Sentences into Four Classes", *Modeling and analysis of information systems*, vol. 29, no. 2, pp. 116-133, 2022.

Нейросетевая классификация русскоязычных предложений по тональности на четыре класса

М. А. Костерин¹, И. В. Парамонов¹

DOI: [10.18255/1818-1015-2022-2-116-133](https://doi.org/10.18255/1818-1015-2022-2-116-133)

¹Ярославский государственный университет им. П. Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.

УДК 004.912

Научная статья

Полный текст на русском языке

Получена 28 апреля 2022 г.

После доработки 23 мая 2022 г.

Принята к публикации 25 мая 2022 г.

Работа посвящена классификации русскоязычных предложений по тональности на четыре класса: положительный, отрицательный, смешанный и нейтральный. В отличие от большинства современных работ в этой области, вводится в рассмотрение класс предложений смешанной тональности. Предложения со смешанной тональностью содержат в себе одновременно и положительно, и отрицательно окрашенную речь.

Для решения данной задачи были применены: нейронная сеть LSTM с механизмом внимания, нейронная сеть GRU с двойным механизмом внимания, нейронная сеть BERT с несколькими модификациями выходного слоя для обеспечения классификации на четыре класса. Эксперименты по сравнению эффективности различных нейронных сетей производилось на трёх корпусах русскоязычных предложений. Два корпуса составлены из пользовательских отзывов: один с отзывами на одежду, другой с отзывами на отели. Третий корпус составлен из новостных статей российских изданий. Лучшая средняя взвешенная F-мера в экспериментах, составляющая 0.90, была достигнута моделью BERT на корпусе отзывов на одежду. На этом же корпусе были отмечены лучшие F-меры для положительных и отрицательных предложений, составившие 0.92 и 0.93 соответственно. Наилучшие показатели классификации нейтральных и смешанных предложений достигаются на корпусе новостных статей. Для них F-мера составляет 0.72 и 0.58 соответственно. В результате экспериментов было продемонстрировано значительное превосходство трансферных нейронных сетей BERT над нейронными сетями предыдущего поколения LSTM и GRU, наиболее ярко выражающееся при классификации текстов со слабо выраженной эмоциональной окраской. Анализ ошибок показал, что на «смежные» классы тональности (положительный/отрицательный и смешанный) приходится большая доля ошибок при классификации с помощью BERT, чем в случае «противоположных» классов (положительный и отрицательный, нейтральный и смешанный).

Ключевые слова: анализ тональности; нейросетевой классификатор; BERT; обработка естественного языка

ИНФОРМАЦИЯ ОБ АВТОРАХ

Максим Алексеевич Костерин | orcid.org/0000-0002-8298-3156. E-mail: makcost@gmail.com
аспирант.

Илья Вячеславович Парамонов | orcid.org/0000-0003-3984-8423. E-mail: ilya.paramonov@fruct.org
автор для корреспонденции | канд. физ.-мат. наук, доцент.

Финансирование: Работа выполнена в рамках инициативной НИР ЯрГУ им. П. Г. Демидова № VIP-016.

Для цитирования: М. А. Kosterin and I. V. Paramonov, "Neural Network-Based Sentiment Classification of Russian Sentences into Four Classes", *Modeling and analysis of information systems*, vol. 29, no. 2, pp. 116-133, 2022.

Введение

В данной работе рассматривается задача классификации предложений на русском языке по тональности на четыре класса: положительный, отрицательный, смешанный и нейтральный. Под задачей классификации тональности в общем виде понимается извлечение и оценка эмоций, вкладываемых автором в текст, независимо от субъекта текста. Предложение называется положительным в том случае, если оно содержит положительные эмоции, факты или выражает положительное мнение автора относительно объекта предложения. Отрицательным называется предложение, которое содержит отрицательные эмоции, факты или выражает отрицательное мнение. Если в предложении присутствуют эмоции или факты из обеих категорий, то такое предложение считается смешанным. Если же предложение не содержит эмоциональной окраски, то оно является нейтральным [1].

Классификация по тональности традиционно применялась к эмоционально экспрессивным текстам, таким как отзывы, для извлечения информации об отношении индивидов или их групп к тому или иному объекту. С появлением более сложных нейронных сетей глубокого обучения, таких как BERT и RoBERTa, классификация тональности стала применяться к более сложным текстам, где тональность предложений выражается менее явно [2].

Данная работа организована следующим образом. Обзор текущего состояния области классификации текстов по тональности приведен в разделе 1. В разделе 2 более подробно рассматривается постановка задачи, а также подходы к её решению. В нём приведено описание различных моделей нейронных сетей, которые применяются в эксперименте. В разделе 3 перечислены используемые корпуса текстов и их характеристики. Результаты применения рассматриваемых моделей к выбранным корпусам приведены в разделе 4.

1. Обзор связанных работ

Задача анализа тональности текста является одной из наиболее распространенных задач обработки естественных языков. Особенно резко важность данной задачи возросла в связи с широким распространением публичных интернет-ресурсов и социальных медиа. Для анализа огромного количества создаваемой пользователями информации были разработаны различные автоматические методы выделения эмоций автора из текста, которые применяются в широком спектре областей от коммерции до политики [3]. Наиболее современными и распространенными из методов анализа тональности на данный момент являются подходы с использованием глубоких нейронных сетей, обзоры которых проводились как для английского [4], так и для русского [5] языков.

Подходы на основе глубокого обучения на текущий момент доминируют в области анализа тональности в связи с двумя ключевыми особенностями: автоматическое выделение признаков рассматриваемых классов тональности и более высокая эффективность. Выделение признаков в сетях глубокого обучения происходит за счет использования множества последовательных слоев нелинейных вычислительных единиц. Высокая эффективность же достигается за счет применения эмбедингов — векторных представлений слов, а также использования графических процессоров в качестве вычислительных устройств. В совокупности это позволяет быстрее выполнять матричные вычисления. Среди подходов на основе глубокого обучения в отдельную группу выделяются методы, использующие трансферное обучение. Их идея заключается в обучении нейронной сети решать одну универсальную задачу на большом объеме данных с последующим применением накопленной в ходе обучения информации для решения специфичных задач. В число трансферных нейронных сетей входят такие модели как OpenAI GPT-2 [6], XLNet [7], BERT [8], имеющие точность классификации тональности на корпусе SST-2 91.8 %, 97.0 %, 94.9 % соответственно.

Глубоким нейронным сетям предшествовали методы машинного обучения, использующие заранее составленные наборы признаков [9]. На основе этих признаков традиционные алгоритмы машинного обучения способны автоматически классифицировать тексты по тональности. Однако

для применения данных методов составление исходных наборов признаков необходимо выполнять вручную, что является сложной и трудоемкой задачей. Вдобавок они показывают результаты хуже, чем модели на основе нейронных сетей. Например, модель из работы [9] достигает точности 87.4 % при классификации тональности твитов.

Существуют также подходы к анализу тональности на основе правил и словарей [10], а не машинного обучения. Их идея заключается в использовании тонального словаря и набора правил лексического и семантического анализа текста для вычисления тональности всего предложения по его содержанию. Преимуществами таких методов является отсутствие необходимости в обучении, что позволяет применять их в случаях, когда нет возможности предоставить большой корпус текстов. Недостатками является существенная сложность построения правил, требующая серьезного языкового моделирования для учёта большого количества способов выражения тональности, свойственных естественному языку.

Задача классификации тональности на 4 класса, решаемая в данной работе, отличается от классической задачи классификации тем, что в ней помимо традиционных классов предложений (положительный, отрицательный, нейтральный) рассматривается также класс предложений смешанной тональности. Предложения со смешанной тональностью содержат в себе одновременно и положительно, и отрицательно окрашенную речь. Данная задача является значительно более редкой в группе задач анализа тональности. Тем не менее её актуальность нельзя недооценивать, так как очень часто тональность реальных предложений невозможно отнести к тому или иному классу однозначно.

Два препятствия встречаются на пути к решению задачи классификации на 4 класса: определение того, что представляет собой смешанная тональность, и подбор корпуса данных, содержащего предложения со смешанной тональностью. Обе проблемы поднимаются в работе [11] применительно к анализу тональности твитов. В ней авторы ставят задачи создания аннотированного множеством рецензентов корпуса твитов и определения наиболее эффективного метода разметки предложений смешанного класса. Для этого они сравнивают два подхода к определению класса предложений смешанной тональности: класс, на котором достигается высокое расхождение в оценках рецензентов, и класс, в котором абсолютное большинство оценок сходится на «смешанной». В эксперименте с использованием первого подхода было достигнуто значение F-меры 0.163 для смешанных предложений, 0.602 — для положительных, 0.786 — для нейтральных, 0.498 — для отрицательных. Значение макро F-меры для эксперимента составило 0.512. При использовании второго подхода значение макро F-меры возросло до 0.803. Важно отметить, что целью экспериментов в работе не ставилось достижение наилучших результатов классификации, а лишь наблюдение за показателями качества при использовании различных выборок. В результате экспериментов делается вывод о том, что второй подход приводит к лучшим результатам классификации.

В работе [12] задача классификации тональности на 4 класса рассматривается в контексте аспектного анализа. Авторы рассматривают смешанную тональность как выражение в одном предложении одновременно положительного и отрицательного мнения об аспекте. Поэтому задачу классификации на 4 класса они сводят к задаче многокомпонентной классификации положительной и отрицательной тональности. Эксперимент с применением данного подхода дал точность 40.38 % для смешанных предложений и 78.50 % для всех классов. Смешанная тональность в рассматриваемой работе выражается как комбинация положительных и отрицательных оценок. Таким образом, определение смешанной тональности само по себе является предметом дискуссий в различных работах.

Задача многоклассовой классификации тональности также имеет и другие варианты. В их число входят достаточно распространенные задачи градуальной классификации тональности и классификации эмоций. Задача градуальной классификации рассматривает набор классов, соответствующих

различной силе выражения тональности, начиная от очень отрицательной и заканчивая очень положительной. Работа [13] посвящена обзору применяющихся при этом подходов и полученных результатов. В ней рассматривается классификация на 5 классов: сильно отрицательный, отрицательный, нейтральный, положительный, сильно положительный. Отличие задачи градуальной классификации от рассматриваемой в настоящей работе классификации на 4 класса заключается в том, что в ней не выделяется смешанная тональность как отдельный класс. Вместо этого предложения, включающие различные эмоции, будут интерпретироваться как нейтральные. Задача классификации эмоций [14] направлена на выделение конкретных эмоций автора из текста, а не его общей тональности. Количество классов в этой задаче также может варьироваться в зависимости от области применения. Данная задача покрывает еще больший спектр эмоциональных окрасок, чем классификация на 4 класса, что делает её более сложной и менее универсальной в плане областей применения.

Таким образом, классификация тональности на 4 класса является отдельной задачей обработки естественных языков, которая обладает своими особенностями и проблемами. Она является менее исследованной, чем традиционная задача анализа тональности, а существующие методы классификации демонстрируют результаты хуже при работе с четырьмя классами тональности. В тоже время, учет класса смешанной эмоционально окраски делает условия задачи более близкими к реальности, в которой тональность предложений не всегда является строго одномерной.

2. Постановка задачи и методы исследования

Целью данной работы является классификация тональности русскоязычных текстов на 4 класса: положительный, отрицательный, нейтральный, смешанный. Методы классификации выбираются из числа современных нейросетевых моделей, продемонстрировавших высокие результаты на различных задачах классификации тональности. Для выбранных подходов вычисляются стандартные характеристики качества на различных корпусах исходных данных, в результате чего делаются выводы об эффективности применяемых подходов.

В данном разделе рассмотрены используемые в настоящей работе модели нейронных сетей.

2.1. Нейронная сеть LSTM с механизмом внимания

До того, как стали широко распространены трансферные нейронные сети на основе трансформеров (transformer), рекуррентные нейронные сети с LSTM (long short-term memory) демонстрировали наиболее высокие результаты в области классификации тональности [15]. В качестве одной из таких нейронных сетей в данной работе рассматривается LSTM с применением функции внимания для назначения соответствующих весов тональным конструкциям в предложении, использование которой применительно к задаче классификации на 3 класса для английского языка демонстрируется в работах [16, 17]. Функция внимания, имеющая вид

$$Attention(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V,$$

принимает в качестве входных данных последовательность X из n -векторных представлений длины d и на основе этой последовательности формирует матрицы весов W_q , W_k , W_v , применяющиеся при вычислении коэффициентов Q , K , V соответственно, где

$$Q = X W_q, \quad K = X W_k, \quad V = X W_v, \quad W_q, W_k, W_v \in \mathbb{R}^{d \times d_k}.$$

Функция внимания для каждого токена вычисляет значение внимания между этим токеном и всеми остальными токенами в предложении. Таким образом выделяется синтаксическая структура, вносящая наибольший вклад в тональность предложения.

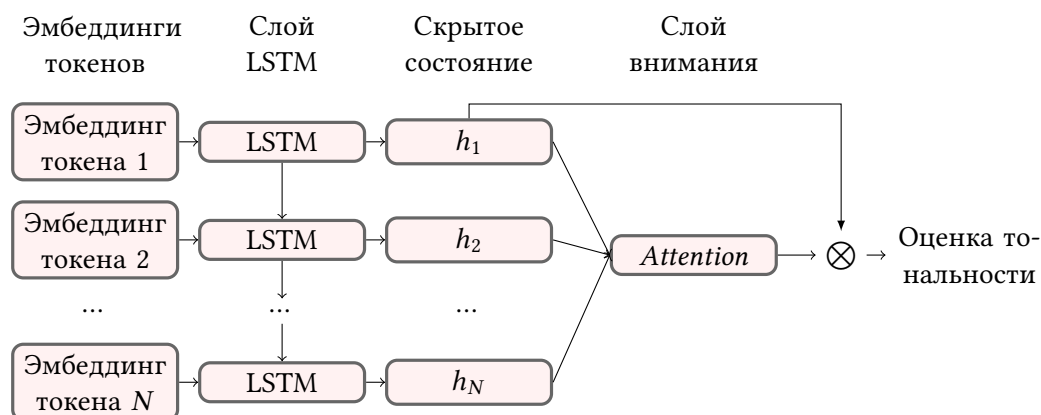


Fig. 1. Attention-based LSTM architecture

Рис. 1. Архитектура сети LSTM с функцией внимания

Схема рассматриваемой в данной работе сети LSTM приведена на рис. 1. В ней выходные данные стандартного слоя LSTM подаются на вход функции внимания, назначение которой заключается в выделении части предложения, вносящей наибольший вклад в эмоциональную окраску всего предложения. Данная архитектура является модификацией модели из работы [16] под задачу классификации тональности текста на 4 класса. В данной работе рассматривалась задача аспектно-ориентированного анализа тональности, поэтому в ходе модификации под рассматриваемую здесь задачу из алгоритма классификации был исключен шаг применения эмбедингов аспектов в вычислении функции внимания. Полученная в результате модификации модель принимает на вход векторные представления токенов предложения, построенные на основе словаря, а результаты выполнения слоя LSTM подает в качестве исходных данных в функцию внимания. Матрица весов, получаемая на выходе слоя внимания, непосредственно применяется при вычислении оценок тональности. В результате работы модели получается оценка для каждого класса, и в качестве тональности предложения выбирается класс с наибольшей оценкой.

2.2. Нейронная сеть GRU с двойным механизмом внимания

В работе [12] для выполнения аспектно-ориентированной классификации тональности текста на 4 класса аспектов предлагается подход с применением сети GRU. Особенность данной модели, по сравнению с другими широко распространенными реализациями LSTM, заключается в использовании двух функций внимания для учета положительной и отрицательной составляющей тональности входной последовательности токенов. Такое архитектурное решение нацелено на улучшение классификации предложений со смешанной тональностью. В предложениях, содержащих как положительно окрашенную, так и отрицательно окрашенную лексику, единственная функция внимания часто формирует оценку всего предложения на основе лишь одной эмоционально окрашенной части предложения, не учитывая другие части, потенциально имеющие противоположную тональность. Механизм двойного внимания призван обрабатывать положительно окрашенные и отрицательно окрашенные части предложения независимо друг от друга, таким образом формируя независимые оценки о содержании эмоционально окрашенной речи в предложении для формирования общей оценки тональности на их основе (рис. 2).

В настоящей работе модель из [12] была модифицирована: слой внимания лишился способности учитывать аспектные эмбединги текста в пользу выполнения классификации тональности всего предложения независимо от его аспектов. Схема сети, получившейся в результате модификации, представлена на рис. 3. В соответствии с этой схемой векторные представления токенов, построенные на основе словаря, проходят через один слой нейронной сети GRU, в результате чего



Fig. 2. Sentiment classification using with double attention

Рис. 2. Определение тональности в модели с механизмом двойного внимания

формируются входные данные для слоя внимания. Обе функции внимания принимают на вход одни и те же данные, но имеют различные матрицы весов и обучаются определять только одну тональность. Выходные данные функций внимания конкатенируются в один вектор, который проходит через сигмоидную функцию активации для получения оценки тональности предложения по двум осям: положительной и отрицательной. На основе этих оценок предложению ставится в соответствие один из четырех классов тональности.

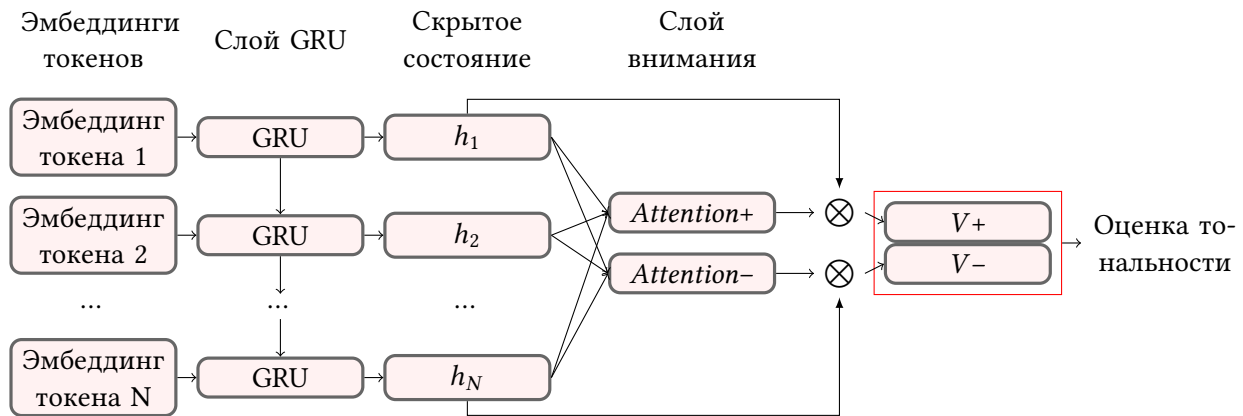


Fig. 3. Dual Attention-based GRU architecture

Рис. 3. Архитектура сети GRU с двойной функцией внимания

2.3. Нейронная сеть BERT

BERT (Bidirectional Encoder Representations from Transformers) — метод предварительной подготовки языковых представлений, обучающий универсальную модель «понимания языка» на большом корпусе неразмеченного текста, а затем использующий эту модель для последующего решения различных задач NLP. Нейронная сеть BERT продемонстрировала рекордные результаты на наборе из 11 основных задач обработки естественных языков, превзойдя языковые модели предыдущего поколения, такие как ELMo и OpenAI GPT [8]. Такие высокие результаты являются следствием того, что BERT является системой глубокого двунаправленного предварительного обучения без учителя, в противовес однонаправленным подходам, применяемым в других моделях.

Архитектура BERT основана на использовании многослойного двунаправленного трансформера, реализация которого базируется на идее, изложенной в работе [18]. Трансформеры для каждого слова строят набор признаков, используя для этого механизм внимания, чтобы выявить «важность» для данного слова всех прочих слов в предложении. Таким образом, построенные признаки для данного слова — сумма линейных преобразований признаков всех слов, взятых в соответствии с их весами «важности».

По своей сути BERT является стеком кодировщиков-трансформеров (encoder transformer), расположенных последовательно таким образом, что передача входных и выходных данных функции

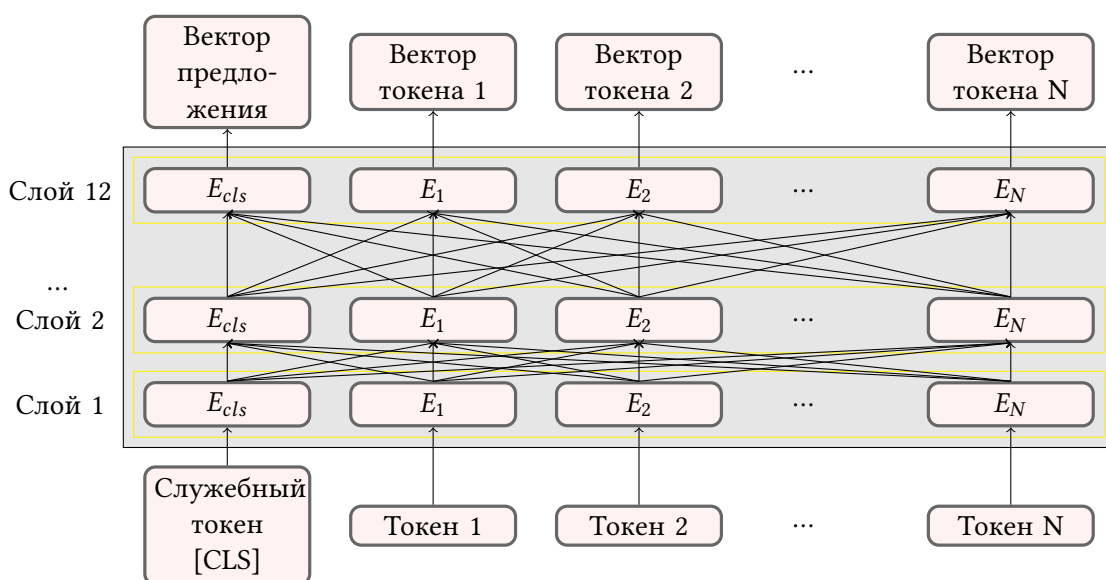


Fig. 4. BERT language model architecture

Рис. 4. Архитектура языковой модели BERT

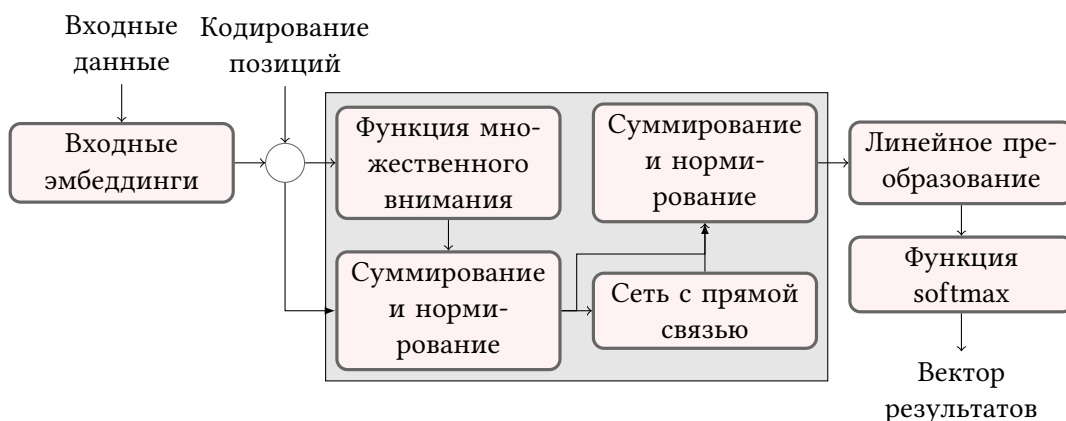


Fig. 5. Operation order in the encoder sublayer of the Transformer architecture

Рис. 5. Последовательность операций в подслое кодировщика архитектуры Transformer

внимания происходит от одного слоя к другому (рис. 4). Количество слоев внутренней реализации BERT варьируется в зависимости от сложности модели. В данной работе применяется базовая модель BERT с количеством слоев $L = 12$ и размером вектора эмбедингов каждой позиции $H = 768$.

Структура кодировщика, использующегося в слоях BERT, представлена на рис. 5. Она состоит из блока функции множественного внимания, сети прямого распространения и двух этапов суммирования и нормирования эмбедингов. Идея заключается в том, что каждое входное слово параллельно проходит через слои кодировщика. При этом на вход в кодировщик передаются не только эмбединги токенов, но и эмбединги позиций слов в предложении. Это делается для того, чтобы устранить недостаток, свойственный архитектуре трансформера, связанный с невозможностью учёта порядка токенов в вычислениях. Данная же модификация позволяет BERT учитывать полный контекст слова во входной последовательности, т. е. одно и то же слово, встречающееся в начале и конце последовательности, будет обрабатываться по-разному.

2.4. Модификации слоя классификации BERT

В своей базовой реализации BERT сначала предобучается «языку» на неразмеченных данных, а затем для адаптации предобученной модели под решение конкретной задачи, в том числе и задачи классификации тональности, требуется постобработка данных, производимых моделью на выходе финального слоя нейронной сети. Данная постобработка производится путем добавления к базовой модели дополнительных слоев, причём даже для одной и той же задачи она может производиться по-разному.

В данной работе рассматриваются три модификации финального слоя нейронной сети BERT для задачи классификации предложений по тональности на 4 класса.

В первой модификации применяется стандартный слой однокомпонентной классификации, используемый авторами модели BERT для оценки её показателей в задаче классификации тональности на 3 класса на корпусе SST-2 [19]. В соответствии с целями данной работы количество классов тональности было увеличено до четырёх путём добавления класса смешанной тональности. Вероятность принадлежности входной последовательности токенов к определенному классу тональности вычисляется посредством умножения вектора предложения, полученного в результате работы модели, на матрицу весов W с последующим применением функции softmax в качестве функции активации и вычисления потерь в соответствии со схемой, приведенной на рис. 6.

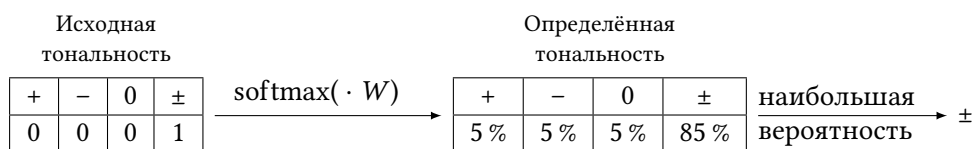


Fig. 6. Single-label sentiment classification with BERT

Рис. 6. Определение тональности в однокомпонентной модели BERT

Вторая модификация использует слой многокомпонентной классификации с компонентами положительной и отрицательной тональности. В отличие от первого подхода, здесь тональность предложения рассматривается не как одно из четырёх взаимоисключающих значений, а как два значения по независимым осям, соответствующим положительной и отрицательной тональностям (рис. 7). В зависимости от соотношения предсказанных моделью вероятностей к определенной совокупности пороговых значений по каждой из осей предложение классифицируется на 4 класса. Вероятность того, что предложение содержит компонент положительной или отрицательной тональности вычисляется посредством умножения вектора предложения, полученного в результате работы модели, на матрицу весов W , с последующим применением сигмоидной функции.

Последняя модификация финального слоя BERT разработана на основе модели D-AT-GRU [12] для задачи аспектного анализа тональности. Оригинальная модель подверглась ряду модификаций с целью адаптации под поставленную в данной работе задачу классификации тональности на 4 класса. Во-первых, подход был адаптирован под задачу классификации тональности предложения

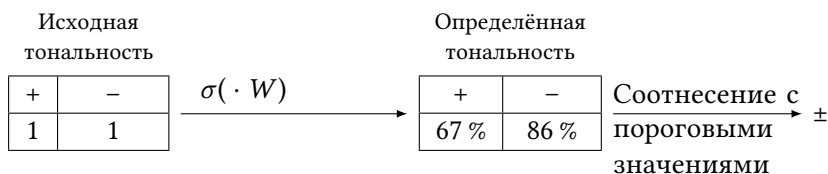


Fig. 7. Multi-label sentiment classification using BERT

Рис. 7. Определение тональности в многокомпонентной модели BERT

в целом в соответствии с описанием, представленным в разделе 2.2. Во-вторых, модель, ранее рассчитанная на применение в сетях GRU, была адаптирована под BERT. В результате проведенных модификаций получился слой BERT, содержащий два независимых компонента, реализующих функцию внимания: один отвечает за классификацию входной последовательности по шкале положительной тональности, а второй — отрицательной. Каждый компонент имеет собственные матрицы весов, таким образом, они являются более специализированными под обнаружение соответствующей составляющей им тональности в предложении. Далее выходы обоих компонентов конкатенируются в двумерный вектор, для которого производится вычисление вероятности нахождения в предложении положительной и отрицательной составляющей тональности с использованием сигмоиды в качестве функции активации. В отличие от ранее приведенных модификаций BERT, данный подход берет из последнего скрытого слоя BERT не выходы уровня всей последовательности токенов, а набор выходных векторов уровня токенов.

3. Корпуса текстов, используемые в эксперименте

Для оценки эффективности рассмотренных нейронных сетей в данной задаче были отобраны три корпуса русскоязычных текстов с различной тематикой и стилем изложения: корпус отзывов на отели, корпус отзывов на одежду, корпус новостных статей из российских СМИ. Включение в набор данных эксперимента корпусов с ярко выраженной тональностью, таких как корпуса отзывов, и корпусов со слабо выраженной тональностью, таких как корпус новостей, нацелено на определение того, насколько сильно выраженность эмоциональной окраски сказывается на результатах классификации тональности с помощью нейронных сетей.

Корпус отзывов на отели формировался на основе отзывов, опубликованных пользователями метапоисковика отелей *trivago*¹. Предварительная разметка тональности корпуса производилась автоматически на основе пользовательской оценки по пятибалльной шкале. Отзывы с оценками «1» и «2» считались отрицательными, с оценками «4» и «5» — положительными, с оценкой «3» — нейтральными. После этапа автоматической разметки выполнялась ручная доработка корпуса группой рецензентов, которые отбирали из корпуса предложения со смешанной тональной окраской и выделяли их в отдельный класс. Сформированный корпус в сумме насчитывает 1205 предложений, среди которых:

- 640 положительных (53.1 %);
- 333 отрицательных (27.6 %);
- 48 смешанных (4.0 %);
- 184 нейтральных (15.3 %).

Корпус отзывов на отели состоит из коротких (табл. 1), часто безличных односоставных предложений. В предложениях в основном описывается узкий набор критериев, характеризующих номер в отеле, таких как: температура, наличие запахов, качество обслуживания и т. д. Вследствие чего лексическое разнообразие является низким и сконцентрированным вокруг специфики предметной области.

Корпус отзывов на одежду является выборкой из корпуса RuReviews [20], который составлен из обзоров в категории «женская одежда и аксессуары», опубликованных покупателями одного из крупных интернет-ресурсов электронной коммерции. Класс тональности каждого отзыва в корпусе RuReviews автоматически определялся на основе пользовательской оценки товара, выставленной по пятибалльной шкале, где оценки «1» и «2» соответствует отрицательной тональности, оценки «4» и «5» — положительной тональности, а оценка «3» — нейтральной. Так как RuReviews включает только 3 класса тональности, группой рецензентов была проведена дополнительная разметка

¹<https://www.trivago.ru>

Table 1. Sentence properties of the hotel reviews corpus**Таблица 1.** Характеристики предложений корпуса отзывов на отели

Характеристика предложения \ Класс тональности	Класс тональности			
	Положительный	Отрицательный	Смешанный	Нейтральный
Минимальная длина	1	1	4	2
Максимальная длина	48	33	24	40
Средняя длина	7	8.8	11.6	7.8
Медиана длины	5	7	12	6

полученной выборки с целью выделения класса смешанной тональности. В результате проведенных работ был сформирован корпус из 2788 предложений, в число которых входят:

- 1065 положительных (38.2 %);
- 1065 отрицательных (38.2 %);
- 540 смешанных (19.4 %);
- 118 нейтральных (4.2 %).

Предложения в данном корпусе в целом являются короткими (табл. 2) и имеют простую синтаксическую структуру. Наличие только одного главного члена предложения характерно для большинства классов. Исключением являются предложения из класса смешанной тональности, средняя длина которых минимум в два раза превосходит среднюю длину предложений других классов, а сами предложения зачастую имеют более одной грамматической основы. Тональность в тех предложениях, где она присутствует, выражена ярко ввиду обилия эмоционально окрашенных слов. Лексика имеет разговорный стиль, что характерно для общедоступной интернет-площадки.

Table 2. Sentence properties of the clothing reviews corpus**Таблица 2.** Характеристики предложений корпуса отзывов на одежду

Характеристика предложения \ Класс тональности	Класс тональности			
	Положительный	Отрицательный	Смешанный	Нейтральный
Минимальная длина	1	1	4	1
Максимальная длина	47	53	90	18
Средняя длина	5	8.3	19.5	5.8
Медиана длины	4	7	15	5

Корпус новостных статей строился на основе выборки предложений из открытого корпуса OpenCorpora². Разметка сформированного набора данных производилась группой экспертов. Предложения корпуса случайным образом распределялись между рецензентами таким образом, чтобы каждое предложение было оценено как минимум тремя участниками группы. При оценке предложению назначался один из пяти классов: положительный, отрицательный, нейтральный, смешанный, неопределимый. Предложения, получившие оценку распределения к классу «неопределимый» хотя бы от одного рецензента, а также предложения, получившие противоположные оценки, отсеивались. Распределение по классам тональности происходило на основе лидирующей оценки среди всех оценок экспертов. Для балансировки классов было отобрано по 500 предложений из каждого класса тональности, в общей сложности формируя корпус из 2000 предложений.

²<http://opencorpora.org>

Table 3. Sentence properties of the Russian news corpus**Таблица 3.** Характеристики предложений корпуса новостных статей

Характеристика предложения	Класс тональности				
		Положительный	Отрицательный	Смешанный	Нейтральный
Минимальная длина		6	6	6	6
Максимальная длина		50	56	92	56
Средняя длина		15.1	16.6	19.6	15.5
Медиана длины		14	15	18	13

Данный корпус имеет наибольшую среднюю длину предложений в каждом классе среди всех рассматриваемых корпусов (табл. 3). Вызвано это, в первую очередь, возросшим смысловым содержанием предложений: увеличилось количество приводимых подробностей, лексика стала более насыщенной. Также доля сложных предложений стала более значительной. Так как в данном корпусе преобладает публицистический стиль речи, эмоциональная окраска стала менее выраженной и не всегда очевидной. Большее лексическое разнообразие данного корпуса делает предложения внутри каждого класса тональности менее схожими между собой.

В качестве дополнительного контрольного корпуса был взят англоязычный корпус отзывов на рестораны из задания 4 воркшопа SemEval-2014. Корпус был разработан для задачи аспектного анализа тональности, но для данной работы он был адаптирован под общий сценарий классификации тональности. Характеристики данного корпуса приведены в табл. 4. Распределение предложений между классами выглядит следующим образом:

- 2454 положительных (57.3 %);
- 897 отрицательных (21.0 %);
- 364 смешанных (8.5 %);
- 565 нейтральных (13.2 %).

Table 4. Sentence properties of the SemEval-2014 restaurant reviews corpus**Таблица 4.** Характеристики предложений корпуса отзывов на рестораны SemEval-2014

Характеристика предложения	Класс тональности				
		Положительный	Отрицательный	Смешанный	Нейтральный
Минимальная длина		1	3	3	1
Максимальная длина		60	69	55	47
Средняя длина		12.8	14.3	18.5	12
Медиана длины		11	13	17	11

4. Результаты экспериментов

Результаты экспериментов для корпусов новостей, отзывов на одежду и отзывов на отели приведены в таблицах 5, 6, 7 соответственно. Используемые метрики включают точность (P), полноту (R) и F -меру (F).

В целом все три модели, основанные на BERT, демонстрируют схожие между собой результаты для предложений с положительной, отрицательной и нейтральной тональностью. Отличие показателей внутри классов тональности в большинстве случаев находится в пределах 5 %. Исключением является эксперимент на корпусе отзывов на одежду, в котором многокомпонентная модификация модели BERT показывает значительно уступающие результаты классификации нейтральных

Table 5. Results obtained via 4-fold cross-validation on the news corpus**Таблица 5.** Результаты четырехкратной кросс-валидации корпуса новостей

Корпус	Полож.			Отриц.			Нейтр.			Смеш.			Взвеш.		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
BERT 1 комп.	0.71	0.69	0.70	0.62	0.64	0.63	0.75	0.71	0.72	0.53	0.55	0.54	0.65	0.65	0.65
BERT 2 комп.	0.73	0.68	0.70	0.64	0.63	0.63	0.72	0.75	0.73	0.55	0.56	0.55	0.66	0.66	0.65
BERT 2 вним.	0.69	0.68	0.68	0.69	0.68	0.68	0.76	0.69	0.72	0.56	0.61	0.58	0.66	0.66	0.66
GRU	0.35	0.32	0.33	0.33	0.32	0.32	0.41	0.43	0.42	0.34	0.37	0.35	0.36	0.36	0.36
LSTM	0.33	0.39	0.35	0.33	0.35	0.34	0.40	0.36	0.38	0.37	0.31	0.33	0.36	0.35	0.35

Table 6. Results obtained via 4-fold cross-validation on the clothing reviews corpus**Таблица 6.** Результаты четырехкратной кросс-валидации корпуса отзывов на одежду

Корпус	Полож.			Отриц.			Нейтр.			Смеш.			Взвеш.		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
BERT 1 комп.	0.93	0.92	0.92	0.93	0.94	0.93	0.73	0.59	0.63	0.50	0.49	0.49	0.90	0.90	0.90
BERT 2 комп.	0.92	0.93	0.93	0.91	0.92	0.91	0.67	0.45	0.52	0.44	0.44	0.43	0.89	0.89	0.89
BERT 2 вним.	0.93	0.92	0.92	0.93	0.92	0.93	0.67	0.65	0.65	0.47	0.54	0.50	0.90	0.90	0.90
GRU	0.91	0.84	0.87	0.85	0.9	0.88	0.25	0.38	0.30	0.33	0.27	0.30	0.84	0.83	0.83
LSTM	0.90	0.86	0.88	0.83	0.91	0.87	0.52	0.33	0.40	0.40	0.28	0.33	0.84	0.84	0.84

Table 7. Results obtained via 4-fold cross-validation on the hotel reviews corpus.**Таблица 7.** Результаты четырехкратной кросс-валидации корпуса отзывов на отели.

Корпус	Полож.			Отриц.			Нейтр.			Смеш.			Взвеш.		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
BERT 1 комп.	0.89	0.91	0.9	0.78	0.83	0.8	0.66	0.68	0.66	0.75	0.08	0.15	0.82	0.82	0.81
BERT 2 комп.	0.90	0.90	0.90	0.82	0.86	0.84	0.66	0.68	0.67	0.46	0.19	0.25	0.82	0.83	0.82
BERT 2 вним.	0.89	0.91	0.90	0.84	0.82	0.83	0.67	0.72	0.69	0.35	0.21	0.25	0.82	0.83	0.82
GRU	0.77	0.78	0.78	0.64	0.52	0.57	0.29	0.38	0.33	0.27	0.25	0.26	0.64	0.63	0.63
LSTM	0.78	0.83	0.80	0.59	0.61	0.60	0.43	0.35	0.39	0.49	0.27	0.35	0.66	0.67	0.66

Table 8. Results obtained via 4-fold cross-validation on the SemEval-2014 restaurant reviews corpus**Таблица 8.** Результаты четырехкратной кросс-валидации корпуса отзывов на рестораны SemEval-2014

Корпус	Полож.			Отриц.			Нейтр.			Смеш.			Взвеш.		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
BERT 1 комп.	0.9	0.91	0.9	0.79	0.79	0.79	0.7	0.62	0.66	0.65	0.66	0.65	0.83	0.83	0.83
BERT 2 комп.	0.9	0.91	0.9	0.79	0.78	0.78	0.63	0.62	0.63	0.64	0.63	0.63	0.82	0.82	0.82
BERT 2 вним.	0.9	0.9	0.9	0.79	0.77	0.78	0.63	0.62	0.62	0.63	0.64	0.63	0.82	0.81	0.81
GRU	0.86	0.86	0.86	0.71	0.66	0.68	0.46	0.48	0.47	0.48	0.49	0.49	0.74	0.74	0.74
LSTM	0.84	0.87	0.85	0.72	0.71	0.71	0.57	0.55	0.56	0.58	0.49	0.53	0.76	0.76	0.76

предложений по сравнению с другими вариантами BERT. При классификации смешанных предложений модель BERT с двойным механизмом внимания показывает несколько лучшие результаты по сравнению с другими трансферными моделями. Данное превосходство больше всего проявляется на наиболее сбалансированном корпусе новостных статей.

Что же касается моделей, основанных на использовании рекуррентных нейронных сетей, то их показатели заметно уступают трансферным моделям. Единственным случаем, в котором они превосходят модели BERT, является классификация предложений смешанной тональности корпуса отзыва на отели, где доля смешанных предложений крайне мала и, видимо, недостаточна для качественного обучения трансферной модели BERT.

При переходе от несбалансированных корпусов с отзывами к сбалансированному корпусу новостей виден резкий прирост качества классификации классов с меньшим количеством предложений. В то же время ввиду роста сложности и неоднозначности предложений снижается качество классификации предложений положительной и отрицательной тональности. При этом снижается влияние вторичных признаков классов, присутствовавших в корпусах отзывов, таких как длина предложения, лексическое разнообразие. Предложения корпуса новостей в этом плане являются более однородными. В корпусах отзывов классы имеют ярко выраженную тональность, чем могут объясняться высокие показатели классификации положительной и отрицательной тональности. В корпусе новостей эмоциональная окраска является более трудноопределимой. На корпусе новостей наилучшие показатели определения тональности присущи предложениям нейтрального класса, предположительно вследствие того, что он не содержит тональных компонентов, свойственных другим классам.

В табл. 8 для сравнения приведены результаты экспериментов для англоязычного корпуса отзывов на рестораны SemEval-2014. На данном корпусе лучше всего себя проявляет стандартная модель BERT с однокомпонентной классификацией. Качество классификации для него уступает качеству классификации на корпусе отзывов на одежду по всем классам, кроме смешанного. Как и остальные корпуса, он уступает корпусу новостей в качестве классификации нейтральных предложений. На основе этого и матрицы ошибок (рис. 8) можно сделать вывод, что класс нейтральных предложений в корпусе с менее выраженной тональностью в целом определяется проще, чем в корпусах с более экспрессивной лексикой.

На рис. 8 приведены матрицы ошибок для экспериментов на корпусе новостных статей. Характерной особенностью, присутствующей в результатах всех моделей BERT, является то, что высокая доля ошибок приходится на класс со схожей составляющей тональности. Например, для класса отрицательной тональности вторая по размеру группа предложений была определена как смешанная. В тоже время доли ошибок с определением отрицательных предложений как смешанных

Fig. 8. Confusion matrices from the news corpora experiment

Actual	Нейтр.	0.71	0.1	0.12	0.07
	Полож.	0.08	0.69	0.06	0.18
	Отриц.	0.1	0.03	0.65	0.23
	Смеш.	0.07	0.16	0.23	0.55
		Нейтр.	Полож.	Отриц.	Смеш.
		Predicted			

(a) BERT 1 комп.

Actual	Нейтр.	0.7	0.13	0.12	0.06
	Полож.	0.07	0.68	0.05	0.21
	Отриц.	0.09	0.05	0.66	0.21
	Смеш.	0.06	0.14	0.19	0.62
		Нейтр.	Полож.	Отриц.	Смеш.
		Predicted			

(c) BERT 2 вним.

Actual	Нейтр.	0.39	0.2	0.19	0.22
	Полож.	0.25	0.35	0.21	0.19
	Отриц.	0.28	0.23	0.36	0.14
	Смеш.	0.26	0.28	0.15	0.31
		Нейтр.	Полож.	Отриц.	Смеш.
		Predicted			

(e) LSTM

Рис. 8. Матрицы ошибок эксперимента на корпусе новостных статей

Нейтр.	0.75	0.08	0.09	0.07
Полож.	0.09	0.69	0.05	0.18
Отриц.	0.13	0.04	0.64	0.21
Смеш.	0.08	0.14	0.22	0.56
	Нейтр.	Полож.	Отриц.	Смеш.
	Predicted			

(b) BERT 2 комп.

Нейтр.	0.32	0.17	0.26	0.25
Полож.	0.2	0.32	0.21	0.28
Отриц.	0.17	0.23	0.43	0.17
Смеш.	0.23	0.25	0.15	0.37
	Нейтр.	Полож.	Отриц.	Смеш.
	Predicted			

(d) GRU

и нейтральных значительно меньше. Такое же поведение наблюдается и для класса положительных и смешанных предложений. Причиной такого поведения является то, что данные классы являются «смежными», т. е. они имеют общие компоненты тональности. У класса нейтральных предложений такой смежности не наблюдается, вследствие чего он имеет меньшую долю ошибок, и, соответственно, лучшие показатели классификации. Для моделей на основе рекуррентных нейронных сетей похожее поведение не наблюдается — они имеют более хаотичное распределение ошибок.

Заключение

В данной работе было проведено сравнение эффективности использования различных моделей нейронных сетей при выполнении классификации тональности на 4 класса. Сравнение производилось на трех корпусах русскоязычных предложений. Два корпуса составлены из пользовательских отзывов, оставленных на Интернет-площадках: один с отзывами на одежду, другой с отзывами на отели. Третий корпус составлен из новостных статей российских изданий. Лучшая средняя взвешенная F-мера экспериментов, составляющая 0.90, была достигнута моделью BERT на корпусе отзывов на одежду. На этом же корпусе были отмечены лучшие F-меры для положительных и отрицательных предложений, составившие 0.92 и 0.93 соответственно. Наилучшие показатели классификации нейтральных и смешанных предложений достигаются на корпусе новостных статей. Для них F-мера составляет 0.72 и 0.58 соответственно. На англоязычном корпусе достигаются лучшие показатели классификации для класса смешанных предложений (F-мера 0.64), однако в остальных классах результаты на русскоязычных корпусах оказываются лучше. В результате экспериментов было продемонстрировано значительное превосходство трансферных нейронных сетей BERT над нейронными сетями предыдущего поколения LSTM и GRU, наиболее ярко выражающееся при классификации текстов со слабовыраженной эмоциональной окраской.

Среди рассмотренных модификаций BERT минимальное превосходство над аналогами показала BERT с механизмом двойного внимания. Оно наиболее выражено при классификации смешанной тональности в сбалансированном корпусе текстов, где прирост F-меры составил 3–4 %. Однако в целом результаты классификации на 4 класса у всех модификаций BERT лежат в пределах статистической погрешности.

Анализ ошибок показал, что на «смежные» классы тональности приходится большая доля ошибок при классификации трансферной нейронной сетью BERT — от 14 до 21 % ошибок, в то время как на «противоположные» классы (положительный и отрицательный, нейтральный и смешанный) всего от 3 до 7 %. Таким образом, дальнейшие исследования в направлении улучшения разграничения «смежных» классов могут положительно сказаться на показателях классификации.

Сравнение результатов классификации на корпусах с сильной и слабой выраженной эмоциональной окраской показало резкое снижение качества классификации положительных и отрицательных предложений при переходе к корпусу с менее выраженной тональностью. Данное снижение особенно заметно для сетей LSTM и GRU, где F-мера упала до 35–36 %. При этом снижении показателей классификации тональных предложений сопутствует рост показателей классификации нейтральных предложений, составивший в среднем 13 % при переходе от корпуса отзывов на одежду к корпусу новостей. Однако однозначный вывод о зависимости качества классификации нейтральных предложений от степени эмоциональной окрашенности корпуса сделать нельзя ввиду дисбаланса классов в используемых корпусах отзывов. Вследствие этого одним из дальнейших направлений исследований видится проведение эксперимента на сбалансированном корпусе предложений с выраженной эмоциональной окрашенностью.

References

- [1] C. Potts, Z. Wu, A. Geiger, and D. Kiela, *Dynasent: A dynamic benchmark for sentiment analysis*, 2020. arXiv: [2012.15349 \[cs.CL\]](#).
- [2] F. Hamborg and K. Donnay, “NewsMTSC: a dataset for (multi-) target-dependent sentiment classification in political news articles”, in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics (ACL), 2021, pp. 1663–1675.
- [3] B. Liu, “Sentiment analysis and opinion mining”, *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.
- [4] O. Habimana, Y. Li, R. Li, X. Gu, and G. Yu, “Sentiment analysis using deep learning approaches: an overview”, *Science China Information Sciences*, vol. 63, no. 1, pp. 1–36, 2020.
- [5] S. Smetanin and M. Komarov, “Deep transfer learning baselines for sentiment analysis in Russian”, *Information Processing & Management*, vol. 58, no. 3, p. 102 484, 2021.
- [6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners”, *Technical report, OpenAI*, 2019.
- [7] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized autoregressive pretraining for language understanding”, *Advances in neural information processing systems*, vol. 32, pp. 5754–5764, 2019.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018. arXiv: [1810.04805v2 \[cs.CL\]](#).
- [9] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, *A convolutional neural network for modelling sentences*, 2014. arXiv: [1404.2188 \[cs.CL\]](#).
- [10] I. Paramonov and A. Poletaev, “Adaptation of Semantic Rule-Based Sentiment Analysis Approach for Russian Language”, in *Proceedings of 30th Conference of Open Innovations Association FRUCT*, IEEE, 2021, pp. 155–164.
- [11] K. Kenyon-Dean, E. Ahmed, S. Fujimoto, J. Georges-Filteau, C. Glasz, B. Kaur, A. Lalande, S. Bhandari, R. Belfer, N. Kanagasabai, *et al.*, “Sentiment analysis: It’s complicated!”, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1886–1895.
- [12] X. Tan, Y. Cai, and C. Zhu, “Recognizing conflict opinions in aspect-level sentiment classification with dual attention networks”, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3426–3431.
- [13] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, “A survey of multimodal sentiment analysis”, *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.
- [14] L. A. M. Oberländer and R. Klinger, “An analysis of annotated corpora for emotion classification in text”, in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2104–2119.
- [15] A. Radford, R. Jozefowicz, and I. Sutskever, *Learning to generate reviews and discovering sentiment*, 2017. arXiv: [1704.01444v2 \[cs.LG\]](#).
- [16] Y. Wang, M. Huang, L. Zhao, and X. Zhu, “Attention-based LSTM for aspect-level sentiment classification”, in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 606–615.

- [17] P. Chen, Z. Sun, L. Bing, and W. Yang, “Recurrent attention network on memory for aspect sentiment analysis”, in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 452–461.
- [18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need”, in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [19] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank”, in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.
- [20] S. Smetanin and M. Komarov, “Sentiment Analysis of Product Reviews in Russian using Convolutional Neural Networks”, in *IEEE 21st Conference on Business Informatics (CBI)*, vol. 1, 2019, pp. 482–486.