Московский государственный университет
имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики

Магистерская программа «Большие данные: инфраструктуры и методы решения задач»

Магистерская диссертация

«Нейросетевой подход к анализу текстов с использованием данных электроокулографии»

Работу выполнил
Федотов Р. А.

Научный руководитель:
кандидат технических наук
Ступников С.А.

Научный консультант:
Шанин И.А.

Москва 2024

Аннотация

Задачи анализа текста находятся на пересечении лингвистики и машинного обучения. Для решения такого класса задач можно использовать данные, полученные путем мультимодальной регистрации. Одним из способов получения таких данных является электроокулография - метод одновременного получения зрительных электрофизиологических сигналов, сопровождающих электроэнцефалографию. При этом выдвигается гипотеза, что упомянутые сигналы функционально зависят как от семантических свойств текста, так и от субъективного отношения человека к этому тексту. В рамках данной работы исследуется, может ли использование таких данных повысить эффективность решений задач предсказания именованных сущностей и анализа тональности текста. Предлагаются подходы к решению этой задачи, описываются модели и приводятся практические результаты, полученные в ходе исследования.

Ключевые слова: электроокулография, электроэнцефалография, нейронная сеть, обработка естественного языка

Содержание

Введение **4**

Обзор родственных работ 7

Цели и задачи работы 9

1 Подход к решению задачи **11**

1.1 Постановка задачи 11

1.2 Метрики оценки моделей 12

1.3 Описание используемых данных 13

1.4 Предобработка текстовых данных 15

1.5 Описание моделей 15

2 Эксперименты и результаты **20**

2.1 Метод обучения и параметры модели 20

2.2 Эксперименты 21

3 Заключение **23**

4 Список литературы **24**

Введение

Обработка текстов (англ. Natural Language Processing, NLP) - это область искусственного интеллекта, занимающаяся созданием компьютерных программ, способных понимать, анализировать и генерировать человеческий язык. Задачи обработки текстов включают в себя множество направлений, таких как машинный перевод [1], распознавание и синтез речи [2], разработка диалоговых систем [3].

Актуальность задач обработки текстов обусловлена несколькими факторами. Во-первых, они позволяют автоматизировать процессы, связанные с обработкой больших объёмов текстовой информации, что особенно важно в условиях современного информационного общества. Во-вторых, они открывают новые возможности для создания интеллектуальных систем, способных взаимодействовать с человеком на естественном языке, что может привести к значительному улучшению качества жизни людей.

Кроме того, задачи обработки текстов имеют широкий спектр применения в различных областях, таких как медицина, юриспруденция, финансы, маркетинг. Они могут использоваться для автоматизации рутинных задач, повышения эффективности работы и принятия решений на основе анализа текстовых данных.

Среди существующих задач можно выделить задачу распознавания именованных сущностей (англ. Named Entity Recognition, NER), целью которой является идентификация и классификация именованных сущностей в тексте. Именованные сущности включают в себя различные объекты, такие как имена людей, названия организаций, географические названия, даты и временные интервалы. Задача NER является важной для многих приложений, включая информационный поиск, обработку запросов, анализ социальных сетей и другие.

Важной задачей также является анализ тональности текста, который направлен на определение эмоциональной окраски текста. Тональность может быть положительной, отрицательной или нейтральной. Алгоритмы, определяющие тональность текста, находят широкое применение в маркетинге для понимания отношения потребителей к продуктам и услугам, а также в мониторинге общественного мнения и анализе со-

циальных медиа.

Для решения данных задач используют глубокие нейронные сети, например, сверточные нейронные сети [4], рекуррентные нейронные сети с долгой краткосрочной памятью[5], сети на основе архитектуры трансформер[6]. Эти модели обучаются на размеченных данных, где каждая сущность в тексте, либо весь текст, имеют уникальный тег. Во время предсказания модель анализирует контекст вокруг каждого слова и присваивает ему соответствующий тег.

Однако можно обрабатывать не только тексты, но и нейрофизиологические измерения, собранные в процессе чтения текста. Новая модальность, включающая в себя данные измерения, может дополнить информацию, содержащуюся в тексте.

Одно из самых популярных и доступных нейрофизиологических измерений на данный момент — электроокулография. Электроокулография регистрирует движения глаз и позволяет получить информацию о том, как люди воспринимают визуальную информацию. По прогнозам исследователей, рынок окулографических устройств растёт на 27,4% ежегодно [7]. Эти устройства просты в настройке и использовании.

Исследования в области анализа человеческого чтения начались в конце 19-го века с работ французского офтальмолога Луи Эмиля Жаваля [8]. Эти исследования не только заложили основу для понимания механизма движений глаз, но и позволили получить ценные сведения о том, как человек читает текст. При помощи стетоскопа, который прикладывался к глазу испытуемого во время чтения напечатанного текста, Жаваль изучил и классифицировал схемы движений глаз. Оказалось, что процесс чтения может быть представлен в виде чередования саккад, то есть быстрых движений глаз для перемещения взгляда из одной точки расположения в другую, и фиксаций, то есть состояний, во время которых глаза в течение определенного промежутка времени остаются неподвижными.

В 20-ом веке исследования советского ученого-физиолога Альфреда Лукьяновича Ярбуса, чьи эксперименты с использованием методов электроокулограммы, расширили представления о движениях глаз и понимании прочитанного [9]. Разработка Робертом С. Кеннеди первой ком-

пьютеризированной системы отслеживания движений глаз в 1960-х годах произвела революцию в изучении поведения при чтении, позволив исследователям проводить высокоточные измерения и анализ движений глаз во время выполнения заданий по чтению. Были использованы данные разных модальностей, полученные помощью электроэнцефалографии (ЭЭГ) и электроокулографии (ЭОГ).

Последующие исследования позволили установить связи между свойствами текста и саккадическими движениями глаз во время чтения. Например, количество и продолжительность фиксаций взгляда на слове зависят от таких лингвистических факторов, как длина слова, частота встречаемости в корпусе и предсказуемость слова в текущем контексте. Влияние частотности слова было выявлено в работе [10] и неоднократно подтверждался в последующих исследованиях [11, 12].

Уровень понимания значения слов, синтаксические категории слов также влияют на продолжительность фиксации [13]. Важно отметить, что даже слова с одинаковой частотой встречаемости могут различаться по степени понимания и предсказуемости слова в контексте [14].

Кроме того, определение релевантности терминов может связана с мозговой активностью в определенных областях мозга [15], также наблюдаются различия в обработке глаголов и существительных, конкретных существительных и абстрактных существительных, а также имен собственных [16].

Различные методы нейровизуализации показали, что частотные диапазоны ЭЭГ соответствуют различным аспектам понимания прочитанного.

Например, тета-активность была связана с когнитивным контролем и рабочей памятью, что крайне важно для обработки и удержания информации во время чтения [17]. Альфа-активность, с другой стороны, связана с внимательностью, указывающей на уровень сосредоточенности и вовлеченности во время чтения [18].

Активность гамма-диапазона использовалась для определения эмоциональных реакций, вызываемых материалом для чтения, проливая свет на аффективные аспекты опыта чтения [19]. Кроме того, бета-частоты играют роль в процессах принятия решений, связанных с осмысленно-

стью текста, влияя на понимание и интерпретацию [20].

Обзор родственных работ

Областью научного исследования в рамках данной магистерской диссертации является решение задач анализа текстовых данных с использованием данных, полученных с помощью мультимодальной регистрации.

Впервые использовать данные электроокулографии в нейросетевых архитектурах начали при решении задач компьютерного зрения [21]. В работе [22] авторы обучают векторные представления окулографических данных для решения задачи классификации изображений; в работе [23] авторы решают задачу детекции объекта на изображении, используя информацию о продолжительности фиксаций при просмотре этого изображения. В данных работах добавление новой модальности в архитектуру происходит путем "ранней интеграции" (англ. early integration), то есть на уровне начальных слоев нейронной сети, не использующих информацию о входном изображении.

Существует множество работ, в которых использование данных мультимодальной регистрации позволило существенно улучшить качество работы моделей в задаче анализа текстовых данных. Так, в работе [24] окулографические данные использовались для решения задачи частеречной разметки (англ. Part-of-speech Tagging, POS). Для этого использовалась bi-LSTM [25] сеть с моделью кодирования слов, которая состоит из иерархической модели, конкатенирующей предварительно обученные векторные представления слов с представлениями подсловсимволов.

В работе [26] использование электроокулографических данных продемонстрировало улучшение производительности модели, решающей задачу суммаризации [27], то есть сокращения длинного текста до более короткой версии, сохраняя при этом его основную идею и ключевые моменты. Признаки, полученные в результате обработки данных о перемещении взгляда, позволили регуляризовать модель.

В работе [28] авторы получают существенное улучшение показателей качества модели в задаче определения ключевых фраз текста. Клю-

чевая фраза состоит из одного или нескольких значимых слов, которые представляют собой основные темы документа. Для этого авторы используют окулографические признаки, полученные из корпуса GECO [29]. В GECO приняли участие носители английского языка, в рамках эксперимента прочитавшие в общей сложности 5031 предложение из произведений классической литературы. Данные, полученные в результате мультимодальной регистрации, в данной модели были использованы в качестве признаков отдельных слов. Для получения векторного представления предложений использовалась bi-LSTM сеть, при этом для интеграции мультимодальных данных использовался механизм внимания [30]. Архитектура модели представлена на Рис. 1.

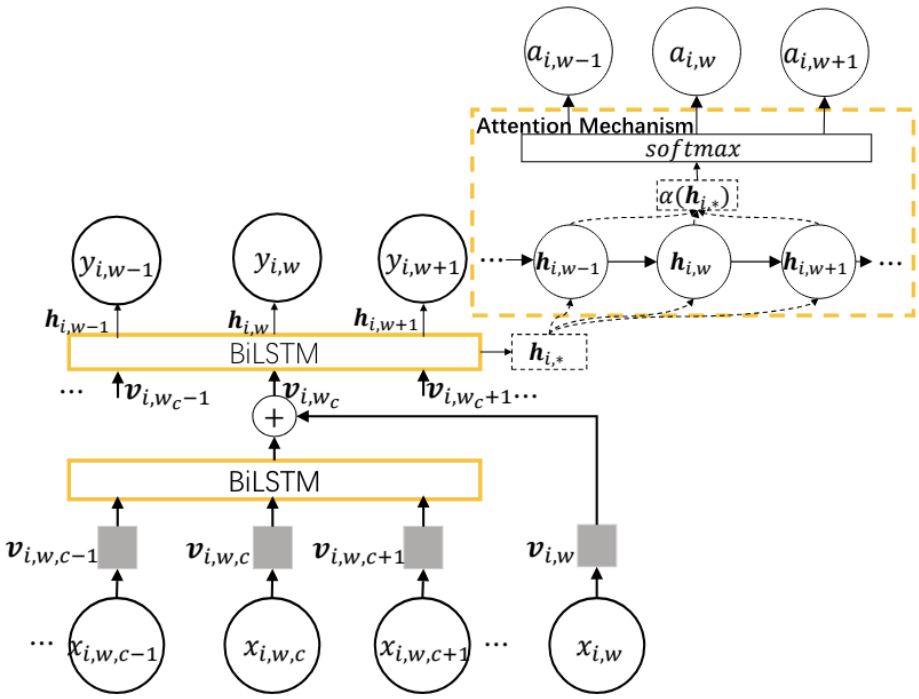

Рис. 1: Схема модели для решения задачи выделения ключевых фраз из документа.

В последние годы в задаче обработки текстовых последовательностей явно доминирующую роль играют модели, основанные на архитектуре трансформер [31], например, большие языковые модели (англ. Large Language Model, LLM)[32]. Важным преимуществом моделей такого класса является возможность одновременной обработки данных разных модальностей.

Например, в работе [33] предлагается модель, работающая одновре-

менно с текстом и изображением. В модели Macaw-LLM [34] рассматривается большая языковая модель, способная обрабатывать изображения, текстовые, аудио- и видео- данные. Для различных модальностей применятся собственные энкодеры, результаты работы которых объединяется и интегрируются в LLM. Архитектура модели представлена на Рис. 2.

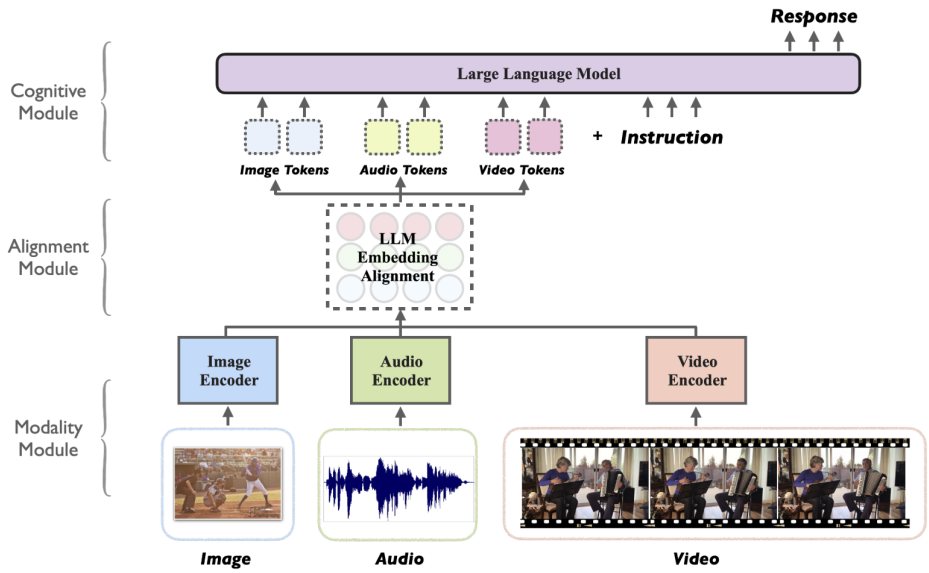

Рис. 2: Схема мультимодальной модели Macaw-LLM.

Исходя из рассмотренной литературы, перспективной является идея использования моделей, основанных на архитектуре трансформер, в решении задач анализа текстовых данных с использованием данных мультимодальной регистрации, таких как ЭЭГ и ЭОГ.

Цели и задачи работы

Целью магистерской диссертации является разработка модели анализа текстовых данных с использованием данных, полученных путем мультимодальной регистрации, на основе архитектуры трансформер.

Для достижения цели необходимо решить следующие задачи:

1. Проведение обзора родственных работ из области машинного обучения;

2. Разработка и применение подходов для предварительного анализа сырых данных;

3. Разработка и применение подходов предобработки текстовых данных для использования в модели машинного обучения;

4. Применение подходящих моделей для решения задач анализа текста с использованием данных мультимодальной регистрации, выбор метрики качества, формирование выборок для кросс-валидации, обучение моделей, подбор гиперпараметров, сравнение моделей по полученным результатам;

5. Анализ ограничений применимости данных, полученных путем мультимодальной регистрации в задачах анализа текстовых данных.

Актуальность работы заключается в создании новой архитектуры обработки текстовых данных с использованием данных мультимодальной регистрации.

1 Подход к решению задачи

1.1 Постановка задачи

Формально опишем задачу, которую необходимо решить: пусть дано предложение $S = (w_1, w_2, ..., w_n)$, где каждому слову w_i поставлен в соответствие набор признаков F_i, соотвествующих данным, полученным в результате мультимодальной регистрации. Для предложения S (или для слова w_i) необходимо определить принадлежность к одному из имеющихся классов.

В качестве частных случаев в данной работе были выбраны две задачи обработки текстовых данных: распознавание именованных сущностей и анализ тональности. Для их решения необходимо построить модели, которые:

1. Для задачи распознования именованных сущностей - определяют пренадлежность каждого слова w_i к одному из 7 классов: BPER, IPER, BORG, IORG, BLOC, ILOC, O (описание классов см. в Таб. 1) [35].

2. Для задачи определения тональности - определяют пренадлежность предложения S к одному из 3 классов: pos (позитивный), neg (негативный), neutral (нейтральный).

Название класса	**Пояснение**
B-PER, I-PER	Слово соответствует началу или находится внутри имени персоны соответственно
B-ORG, I-ORG	Слово соответствует началу или находится внутри названия организации соответственно.
B-LOC, I-LOC	Слово соответствует началу или находится внутри названия локации.

Таблица 1: Используемые классы в задаче распознавания именованных сущностей (NER)

Ключевым отличием задачи анализа тональности от задачи распознавания именованных сущностей является тот факт, что тональность

текста является субъективным показателем, который зависит от читающего текст и от его контекста. В отличие от задачи распознавания именованных сущностей, невозможно в общем виде дать лингвистическую классификацию предложений по уровню тональности [36].

1.2 Метрики оценки моделей

Так как обе описанных выше задачи являются задачами классификации, то для их оценки можно использовать стандартные метрики классификации, как precision, recall и F1.

Взвешенная точность (англ. precision) - это метрика классификации, которая измеряет среднюю точность по всем классам, при этом каждый класс взвешивается по степени его встречаемости в выборке. В контексте NER взвешенная точность учитывает дисбаланс классов в наборе данных для каждого типа именованных сущностей. Аналогично, в задече определения тональности взвешенная точность учитывает распределение каждой категории в имеющихся данных. Взвешенная точность может быть опредена следующим образом:

$$\text{precision}_{weighted} = \frac{\sum_{i=1}^{N} \text{precision}_i \times \text{support}_i}{\sum_{i=1}^{N} \text{support}_i},$$

где precision_i есть precision класса i, support_i есть число сущностей класса i в датасете, N есть общее число классов.

Полностью аналогично данной формуле можно ввести понятия взвешенной полноты (англ. recall) и взвешенного F1:

$$\text{recall}_{weighted} = \frac{\sum_{i=1}^{N} \text{recall}_i \times \text{support}_i}{\sum_{i=1}^{N} \text{support}_i},$$

$$\text{F1}_{weighted} = \frac{\sum_{i=1}^{N} \text{F1}_i \times \text{support}_i}{\sum_{i=1}^{N} \text{support}_i},$$

где recall_i есть recall класса i, F1_i есть F1 класса i.

Взвешенные точность, полнота и F1 учитывают дисбаланс классов и дают комплексную оценку эффективности моделей. В задаче распознования именованных сущностей важно учитывать данные метрики для

каждого класса именованных сущностей, для учета дисбаланса классов в обучающей и тестовой выборке. Аналогично, при анализе тональности оценка эффективности для различных категорий настроений имеет решающее значение, поскольку распределение тональностей в генеральной совокупности может быть различным. Взвешенные метрики гарантируют, что оценка отражает общую эффективность модели с учетом распределения классов в данных.

1.3 Описание используемых данных

Цюрихский корпус когнитивной обработки языка (англ. The Zurich Cognitive Language Processing Corpus, [37]) является основным источником данных для данной работы. Это первый свободно доступный набор данных мультимодальной регистрации с одновременной записью трекинга глаз и ЭЭГ при чтении предложений. Этот корпус включает записи 12 взрослых носителей языка, читающих около 1100 английских предложений.

Корпус содержит как записи естественного чтения, так и записи чтения с решением задач. В данной работе используется запись данных естественного чтения, в которых испытуемые читали с собственной скоростью и без какого-либо конкретного задания, кроме ответов на контрольные вопросы, проверяющие понимание прочитанного.

Часть предложений, предназначенная для решения задачи распознавания именованных сущностей, включает в себя 647 предложений (7737 токенов) из статей Википедии [38]. Вторая часть предложений содержит 400 положительных, отрицательных и нейтральных предложений (8138 токенов) из корпуса Stanford Sentiment Treebank [39], для анализа эмоций и мнений во время чтения. Все 12 испытуемых читали одни и те же предложения.

Набор мультимодальных данных состоит из следующих признаков [35], соответствующих регистрации данных движения глаз:

- количество фиксаций (NFIX);
- количество всех фиксаций, приходящихся на слово;

- продолжительность первой фиксации (FFD);
- продолжительность первой фиксации на текущем слове;
- общее время чтения (TRT);
- сумма всех продолжительностей фиксаций на текущем слове;
- продолжительность взгляда (GD) - сумма всех фиксаций на текущем слове при чтении первого прохода до того, как взгляд переместится за пределы слова;

Для повышения устойчивости признаков и уменьшения дисперсии данные усреднялись по всем испытуемым. Поскольку ЭОГ- и ЭЭГ-характеристики записывались одновременно, авторы ZuCo смогли извлечь ЭЭГ-характеристики на уровне слов. Во время предварительной обработки ZuCo использовали 23 электрода в наружной части окружности головы (подбородок и шея). Таким образом, каждый признак ЭЭГ, соответствующий длительности конкретной фиксации, содержит 105 показателей, полученных от электродов.

Сигнал ЭЭГ разделен на 8 частотных диапазонов, которые представляют собой фиксированные диапазоны частот и амплитуд волн на временной шкале. Данные о интервальном кодировании приведены в Таб. 2.

Частотный диапазон	**Частотность (в Гц)**
тета 1	от 4 до 6
тета 2	от 6.5 до 8
альфа 1	от 8.5 до 10
альфа 2	от 10.5 до 13
бета 1	от 13.5 до 18
бета 2	от 13.5 до 18
гамма 1	от 18.5 до 30
гамма 2	от 40 до 49.5

Таблица 2: Частотные диапазоны, используемые в интервальном кодировании

1.4 Предобработка текстовых данных

Из предложений, содержащихся в наборе данных, были удалены специальные символы, знаки препинания. Слова были переведены в нижний регистр, дополнительно удалены распространенные стоп-слова (например, "and", "the", "is"), которые не несут существенной смысловой нагрузки и могут вносить шум в модель. Дополнительно проведена лемматизация, то есть сокращение слов до их базовой или корневой формы для нормализации текста и уменьшения разреженности в пространстве признаков.

После выполнения данного этапа предобработки изучаемый набор данных будет состоять исключительно из фиксаций и саккад, относящиеся к процессу чтения.

1.5 Описание моделей

В данной работе исследованы два вида моделей. В качестве базовой модели была выбрана модель [35], которая использует рекуррентные нейронные сети для агрегации данных мультимодальной регистрации. Векторные представления слов извлекаются из предобученной модели word2vec [40].

В качестве исследуемой модели была выбрана модель, которая использует механизм внимания для агрегации данных мультимодальной регистрации. В качестве текстовых эмбеддингов используется выход модели энкодера BERT [41].

Для решения задачи распознавания именованных сущностей в каждой модели построен классификатор с 7 классами, применяемый к каждому слову в предложении. Для решения задачи анализа тональности предложения в каждой модели построен классификатор с 3 классами, использующий вектор последнего слова в предложении, который после выхода из энкодера содержит информацию о предыдущих словах.

Архитектура базовой модели может быть определена следующим образом:

$$
\begin{aligned}
\text{Слой эмбеддингов слов:} \quad & \mathbf{X}_w = \text{Word2Vec}(w_1, w_2, ..., w_n) \\
\text{Категорийный энкодер:} \quad & \mathbf{H}_f = \text{LSTM}_{\text{Cat}}(F_1, F_2, ..., F_n) \\
\text{Конкатенация:} \quad & \mathbf{X} = [\mathbf{X}_w, \mathbf{H}_f] \\
\text{Пословный энкодер:} \quad & \mathbf{H} = \text{LSTM}_{\text{Word}}(\mathbf{X}) \\
\text{NER слой:} \quad & \hat{y}_{\text{NER}} = \text{softmax}(\mathbf{W}_{\text{NER}}\mathbf{H} + \mathbf{b}_{\text{NER}}) \\
\text{Sentiment Analysis слой:} \quad & \hat{y}_{\text{SA}} = \text{softmax}(\mathbf{W}_{\text{SA}}\mathbf{H} + \mathbf{b}_{\text{SA}})
\end{aligned}
$$

при этом:

- $\mathbf{X}_w$ - векторы, составленные из эмбеддингов слов модели word2vec $w_1, w_2, ..., w_n$.
- $\mathbf{H}_f$ - векторы, полученные от слоя категорийной агрегации, отражающая контекстную информацию категориальных признаков $F_1, F_2, ..., F_n$.
- $\mathbf{X}$ - конкатенация эмбеддингов слов и агрегированных категориальных признаков.
- $\mathbf{H}$ - векторы, отражающие контекстную информацию входной последовательности.
- $\hat{y}_{\text{NER}}$ прогнозируемый результат для задачи распознавания именованных сущностей с использованием функции активации softmax.
- $\hat{y}_{\text{SA}}$ - прогнозируемый результат для задачи Sentiment Analysis с использованием функции активации softmax.
- $\mathbf{W}_{\text{NER}}$ и $\mathbf{W}_{\text{SA}}$ - матрицы для задач NER и SA, соответственно.
- $\mathbf{b}_{\text{NER}}$ и $\mathbf{b}_{\text{SA}}$ - векторы смещения для задач NER и SA, соответственно.

У моделей, использующих рекуррентные нейронные сети, есть ряд известных недостатков: LSTM плохо улавливают долгосрочные зависимости в последовательностях, что приводит к снижению производительности по мере увеличения длины последовательности [42]; при обучении

таких часто возникает проблема "исчезающего" градиента [43]; LSTM склонны к переобучению на небольших выборках [44].

В данной работе предлагается следующая модель, использующая архитектуру трансформер в своей основе.

$$
\begin{aligned}
\text{Слой эмбеддингов слов:} \quad & \mathbf{X}_w = \text{BERT}(w_1, w_2, ..., w_n) \\
\text{Энкодер ЭЭГ сигналов:} \quad & \mathbf{X}_e = \text{MultiHeadAttention}_{\text{EEG}}(e_1, e_2, ..., e_n) \\
\text{Энкодер айтрекинг сигналов:} \quad & \mathbf{X}_t = \text{MultiHeadAttention}_{\text{ET}}(t_1, t_2, ..., t_n) \\
\text{Конкатенация:} \quad & \mathbf{X} = [\mathbf{X}_w, \mathbf{X}_e, \mathbf{X}_t] \\
\text{Позиционное кодирование:} \quad & \mathbf{X}_{\text{pos}} = \text{PositionalEncoding}(\mathbf{X}) \\
\text{Слой агрегации MHSA 1:} \quad & \mathbf{H}_1 = \text{MultiHeadAttention}_1(\mathbf{X}_{\text{pos}}) \\
\text{Линейный слой 1:} \quad & \mathbf{Z}_1 = \text{FeedForward}_1(\mathbf{H}_1) \\
\text{Слой агрегации MHSA 2:} \quad & \mathbf{H}_2 = \text{MultiHeadAttention}_2(\mathbf{Z}_1) \\
\text{Линейный слой 2 2:} \quad & \mathbf{Z}_2 = \text{FeedForward}_2(\mathbf{H}_2) \\
\text{NER слой:} \quad & \hat{y}_{\text{NER}} = \text{Softmax}(\mathbf{W}_{\text{NER}}\mathbf{Z}_2 + \mathbf{b}_{\text{NER}}) \\
\text{Sentiment Analysis слой:} \quad & \hat{y}_{\text{SA}} = \text{Softmax}(\mathbf{W}_{\text{SA}}\mathbf{Z}_2 + \mathbf{b}_{\text{Mood}})
\end{aligned}
$$

при этом:

- $\mathbf{X}_w$ - векторы, полученные путем энкодинга моделью BERT слов $w_1, w_2, ..., w_n$.
- $\mathbf{X}_e$ - векторы, полученные из эмбеддингов категорийных признаков ЭЭГ $e_1, e_2, ..., e_n$.
- $\mathbf{X}_t$ - векторы, полученные из эмбеддингов категорийных признаков айтрек данных $t_1, t_2, ..., t_n$.
- $\mathbf{X}$ - векторы, полученные конкатенацией векторов $\mathbf{X}_w, \mathbf{X}_e, \mathbf{X}_t$
- $\mathbf{X}_{\text{pos}}$ - векторы позиционного кодирования $\mathbf{X}$.
- $\mathbf{H}_1$ и $\mathbf{H}_2$ - векторы, полученные после экодинга 1 и 2 MHSA слоем.
- $\mathbf{Z}_1$ и $\mathbf{Z}_2$ - векторы, полученные после 1 и 2 линейного слоя.

- $\hat{y}_{\text{NER}}$ - прогнозируемый результат для задачи NER с использованием функции активации softmax.
- $\hat{y}_{\text{Mood}}$ - прогнозируемый результат для задачи Sentiment Analysis с использованием функции активации softmax.
- $\mathbf{W}_{\text{NER}}$ и $\mathbf{W}_{\text{SA}}$ - матрицы для задач NER и SA, соответственно.
- $\mathbf{b}_{\text{NER}}$ и $\mathbf{b}_{\text{SA}}$ - векторы смещения для задач NER и SA, соответственно.

Схема модели представлена на Рис. 2. Эта модель использует механизм Multi-Head Self-Attention (MHSA) в архитектуре Transformer для улавливания глобальных зависимостей и отношений во входных данных. Компонент MHSA реализован в соответствии с оригинальной формулировкой, представленной в [45] , где несколько голов внимания одновременно используют различные части входной последовательности, что повышает способность модели посещать релевантную информацию в различных позициях последовательности.

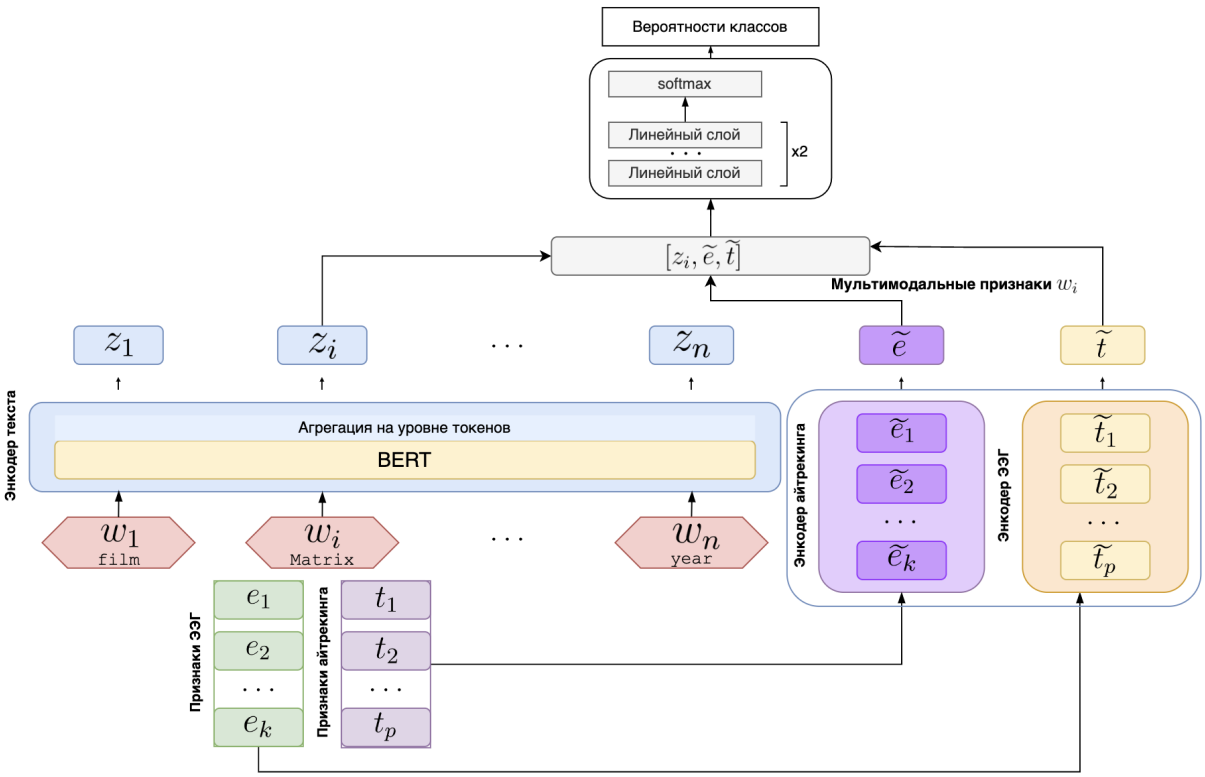

Рис. 3: Схема модели для решения задачи классификации именованных сущностей и анализа тональности.

Векторы позиционного энкодинга $\mathbf{X}_{\text{pos}}$ в соответствии с работой [46] вычисляются следующим образом:

$$\mathbf{X}_{\text{pos}(i,2j)} = \sin\left(\frac{i}{10000^{2j/d}}\right)$$

$$\mathbf{X}_{\text{pos}(i,2j+1)} = \cos\left(\frac{i}{10000^{2j/d}}\right)$$

где i, j есть координаты в матрице $\mathbf{X}_{\text{pos}}$, d - размерность эмбеддингов.

2 Эксперименты и результаты

2.1 Метод обучения и параметры модели

Все модели были обучены на описанном в ч. 1 датасете методом 5-fold кросс-валидации (с разделением в пропорциях - 80% обучение, 10% валидация, 10% тест), а ранняя остановка осуществлялась после 5 эпох ухудшения метрик на валидационной выборке.

Для инициализации весов эмбеддингов слов для бейзлайн-модели использовались предобученные эмбеддинги из библиотеки $gensim$, набор "word2vec-google-news-300".

В качестве энкодера последовательности слов в модели использовалась предобученная модель BertModel из библиотеки $transformers$ "bert-base-uncased". Для инициализации остальных весов модели использовалось нормальное Ксавьер распределение [47]. Размерность обучаемых эмбеддингов равна $d = 128$.

Бейзлайн-модели обучались с помощью модифицированного метода стохастического градиентного спуска Adam [48] с параметрами $\gamma = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1\text{e}{-8}$.

Модели на основе механизма внимания обучались с помощью модифицированного метода стохастического градиентного спуска AdamW [48] с параметрами $\gamma = 0.002, \beta_1 = 0.5, \beta_2 = 0.98, \epsilon = 1\text{e}{-8}, \lambda = 0.01$.

Так как в обоих случаях решалась задача классификации с диспропорциями распределения классов в данных, в качестве функции потерь была выбрана взвешенная многоклассовая кросс-энтропия. Пусть N - размер выборки, C число классов, y_{ij} - метка объекта i для класса j, p_{ij} - предсказанная вероятность для объекта i относится к классу j, и s_j размер класса j.

Введем вес класса j w_j как отношение вида

$$w_j = \frac{N}{s_j \cdot C},$$

тогда функция потерь может быть определена как:

$$CE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} w_j \cdot y_{ij}\log(p_{ij}).$$

2.2 Эксперименты

Разработка методов и подходов, предложенных в рамках исследовании магистерской диссертации, проводилась на языке *Python* на платформе *Google Colaboratory*. Применялись следующие пакеты библиотек:

- *pytorch* - для описания модели архитектуры модели и обучения использовался .
- *pandas* и *numpy* – при работе с данными в виде таблиц (датафреймов);
- *pickle* – для сериализации данных обучения, валидации и тестирования модели;
- *sklearn* – для подсчета метрик точности моделей и k-fold кросс-валидации;

В Таб. 3 представлены результаты проведенных в рамках данной работы экспериментов. В данной таблице "baseline" - базовая модель на основе векторов word2vec без признаков мультимодальной регистрации, "baseline + EEG + gaze" - базовая модель с признаками мультимодальной регистрации (ЭЭГ и координаты движений глаз), "attn" - модель на основе трансформера без признаков мультимодальной регистрации, модель "attn+EEG" - модель на основе трансформера с использованием признаков ЭЭГ данных, модель "attn+gaze" - модель на основе трансформера с ипользованием признаков координат движения глаз, модель "attn+EEG+gaze" - модель на основе трансформера с использованием признаков координат движения глаз и ЭЭГ. Используются метрики P (precision), R (recall), F1.

Изучив полученные результаты можно прийти к следующим выводам:

1. У моделей на основе архитектуры трансформер лучшая способность обучаться и извлекать дополнительную информацию из новых данных. Добавление данных мультимодальной регистрации в задаче

Модель	NER			Sentiment Analysis		
	P	R	F1	P	R	F1
baseline	0.64	0.77	0.69	0.78	0.79	0.78
baseline+EEG+gaze	0.66	0.79	0.72	0.81	0.81	0.81
attn	0.66	0.80	0.71	0.82	0.85	0.84
attn+EEG	0.67	0.81	0.72	0.83	0.85	0.84
attn+gaze	0.69	0.81	0.74	0.88	0.86	0.87
attn+EEG+gaze	0.70	0.82	0.75	0.89	0.88	0.88

Таблица 3: Сравнение результатов различных моделей в задаче распознавания именованных сущностей и в задаче определения тональности.

определения тональности для модели, основанной на LSTM, позволяет получить 3.85% прирост в точности. При этом добавление тех же данных в модель на основе архитектуры трансформер позволяет получить прирост на 8.54%.

3. Данные мультимодальной регистрации значительно улучшают качество работы базовой модели и модели на основе трансформеров в задаче анализа тональности. Это может быть объяснено более простой структурой задачи и субъективностью оценки тональности текста[49].

4. Подтверждается тот факт [35], что при усреднении по испытуемым признаки, построенные на ЭЭГ данных не дают существенного улучшения в задаче анализа текстов.

5. Совместное использование ЭЭГ данных и координат движения глаз во время чтения дает больший эффект для нейросетевых моделей на основе трансформера. Это также может быть объяснено большей сложностью модели и способностью к обучению на небольшом наборе данных[45].

3 Заключение

Электроокулография — один из ключевых инструментов в изучении человеческого сознания. Данные, полученные с помощью этой технологии, помогают найти ответы на многие вопросы о работе мозга. В рамках исследования изучалась возможность разработки модели анализа текстовых данных и данных мультимодальной регистрации. Собраны признаки, относящиеся к различным типам мультимодальной регистрации - частотные интервалы электроэнцефалограммы (ЭЭГ) и признаки координат движения глаз. На основе этих данных обучены и протестированы модели на основе рекуррентных нейронных сетей и на основе трансформера с механизмом внимания. Сделаны содержательные выводы о свойствах и особенностях этих моделей, выделены ограничения применимости данных мультимодальной регистрации в решении текстовых задач.

Автор благодарит С.А.Ступникова, И.А.Шанина за ценные замечания и за помощь в работе над исследованием.

4 Список литературы

[1] Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. A survey of deep learning techniques for neural machine translation, 2020.

[2] Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai, 2023.

[3] Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. Recent advances in deep learning based dialogue systems: A systematic survey, 2022.

[4] Zewen Li, Wenjie Yang, Shouheng Peng, and Fan Liu. A survey of convolutional neural networks: Analysis, applications, and prospects, 2020.

[5] Benyamin Ghojogh and Ali Ghodsi. Recurrent neural networks and long short-term memory networks: Tutorial and survey, 2023.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[7] Yili Jin, Junhua Liu, Fangxin Wang, and Shuguang Cui. Where are you looking?: A large-scale dataset of head and gaze behavior for 360-degree videos and a pilot study. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22. ACM, October 2022.

[8] A. I. Medvedeva and M. V. Kholod. Using artificial intelligence methods for the studyed visual analyzer, 2024.

[9] A. I. Medvedeva and M. V. Kholod. Artificial intelligence methods, 2022.

[10] Clara Meister, Tiago Pimentel, Thomas Hikaru Clark, Ryan Cotterell, and Roger Levy. Analyzing wrap-up effects through an information-theoretic lens, 2024.

[11] Laura Wall, David Gunawan, Scott D. Brown, Minh-Ngoc Tran, Robert Kohn, and Guy E. Hawkins. Identifying relationships between cognitive processes across tasks, contexts, and time, 2020.

[12] Rohit Voleti, Julie M. Liss, and Visar Berisha. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):282–298, February 2020.

[13] Albert Le Floch and Guy Ropars. Hebbian control of fixations in a dyslexic reader, 2023.

[14] Jon Gauthier and Roger Levy. The neural dynamics of auditory word recognition and integration, 2023.

[15] Ruotsalo T. Spapé M.M. Kosunen I. Barral O. Ravaja N. Kaski S. Eugster, M.J.A. Predicting term-relevance from brain signals, 2014.

[16] Mueller M. Weiss S. The contribution of eeg coherence to the investigation of language, 2003.

[17] Kelly Williams and Sharon Vaughn. Williams vaughn (2019) effects of an intensive reading intervention for ninth-grade english learners with learning disabilities. *Learning Disability Quarterly*, 43, 06 2019.

[18] Wolfgang Klimesch. ??-band oscillations, attention, and controlled access to stored information. *Trends in cognitive sciences*, 16, 11 2012.

[19] Mu Li and Bao-Liang Lu. Emotion classification based on gamma-band eeg. *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2009:1323–6, 09 2009.

[20] Manuel Eugster, Tuukka Ruotsalo, Michiel Spapé, Ilkka Kosunen, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. Predicting term-relevance from brain signals. 07 2014.

[21] Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. Improving natural language processing tasks with human gaze-guided neural attention, 10 2020.

[22] Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. Gaze embeddings for zero-shot image classification, 2017.

[23] Iaroslav Shcherbatyi, Andreas Bulling, and Mario Fritz. Gazedpm: Early integration of gaze information in deformable part models, 2015.

[24] Sigrid Klerke and Barbara Plank. At a glance: The impact of gaze aggregation views on syntactic tagging. In Aditya Mogadala, Dietrich Klakow, Sandro Pezzelle, and Marie-Francine Moens, editors, *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 51–61, Hong Kong, China, November 2019. Association for Computational Linguistics.

[25] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging, 2015.

[26] Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. Improving sentence compression by learning to predict gaze. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California, June 2016. Association for Computational Linguistics.

[27] Divakar Yadav, Jalpa Desai, and Arun Kumar Yadav. Automatic text summarization methods: A comprehensive review, 2022.

[28] Yingyi Zhang and Chengzhi Zhang. Using human attention to extract keyphrase from microblog post. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5867–5872, Florence, Italy, July 2019. Association for Computational Linguistics.

[29] Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49, 05 2016.

[30] Saidul Islam, Hanae Elmekki, Ahmed Elsebai, Jamal Bentahar, Najat Drawel, Gaith Rjoub, and Witold Pedrycz. A comprehensive survey on applications of transformers for deep learning tasks, 2023.

[31] Xiangyang Mou, Mo Yu, Bingsheng Yao, and Lifu Huang. Efficient long sequence encoding via synchronization, 2022.

[32] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[34] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration, 2023.

[35] Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. Advancing nlp with cognitive language processing signals, 2019.

[36] Erik McGuire and Noriko Tomuro. Sentiment analysis with cognitive attention supervision. *Proceedings of the Canadian Conference on Artificial Intelligence*, 06 2021.

[37] Nora Hollenstein, Jonathan Rotsztejn, Marius Tröndle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific Data*, 5:180291, 12 2018.

[38] Aron Culotta, Andrew Mccallum, and Jonathan Betz. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. 01 2006.

[39] Richard Socher, A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. *EMNLP*, 1631:1631–1642, 01 2013.

[40] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[42] Larkin Liu, Yu-Chung Lin, and Joshua Reid. Improving the performance of the lstm and hmm model via hybridization, 2021.

[43] Phong Le and Willem Zuidema. Quantifying the vanishing gradient and long distance dependency problem in recursive neural networks and recursive lstms, 2016.

[44] Aysu Ezen-Can. A comparison of lstm and bert for small corpus, 2020.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[46] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers, 2023.

[47] Justin Sirignano and Konstantinos Spiliopoulos. Scaling limit of neural networks with the xavier initialization and convergence to a global minimum, 07 2019.

[48] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[49] Fatma Jemai, Mohamed Hayouni, and Sahbi Baccar. Sentiment analysis using machine learning algorithms. pages 775–779, 06 2021.