Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет
имени первого Президента России Б. Н. Ельцина»
Институт радиоэлектроники и информационных технологий-РТФ
Кафедра информационных технологий и систем управления

ДОПУСТИТЬ К ЗАЩИТЕ ПЕРЕД ГЭК
Заведующий кафедрой ИТиСУ
_____ Е. В. Кислицын
«30» мая 2025 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Краулер депрессивного поведения в социальных сетях

Научный руководитель: Сорокин Артем Константинович, к.т.н, доцент
ученая степень, ученое звание подпись

Нормоконтролер: Огуренко Егор Владимирович подпись

Студент группы: РИМ-230962 Ефимович Евгений Александрович подпись

Екатеринбург
2025

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Уральский федеральный университет имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий – РТФ
Кафедра информационных технологий и систем управления
Направление подготовки 09.04.01 Информатика и вычислительная техника
Образовательная программа 09.04.01/33.03 Инженерия машинного обучения

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студента Ефимович Евгений Александрович **группы** РИМ-230962
(фамилия, имя, отчество)

1. Тема выпускной квалификационной работы
Краулер депрессивного поведения в социальных сетях
Утверждена распоряжением по институту от «02» декабря 2025 г. № 33.02-05/334

2. Научный руководитель
Сорокин Артем Константинович, к.т.н, доцент

3. Исходные данные к работе
Нормативная, учебная, методическая литература по теме магистерской диссертации, материалы, полученные в ходе преддипломной практики

4. Перечень демонстрационных материалов
Презентация, архитектура программного комплекса, приложение

5. Календарный план

№ п/п	Наименование этапов выполнения работы	Срок выполнения этапов работы	Отметка о выполнении
1.	1 раздел (глава)	до 24.03.2025 г.	Выполнено
2.	2 раздел (глава)	до 28.04.2025 г.	Выполнено
3.	3–4 раздел (глава)	до 19.05.2025 г.	Выполнено
4.	ВКР в целом	до 23.05.2025 г.	Выполнено

Научный руководитель Сорокин Артем Константинович _____ (подпись)
Ф.И.О.

Студент задание принял к исполнению 10.02.2025 г. _____ (подпись)
дата

6. Допустить Ефимович Евгения Александровича к защите выпускной квалификационной работы в экзаменационной комиссии

Заведующий кафедрой ИТиСУ _____ Е. В. Кислицын
(подпись) Ф.И.О.

РЕФЕРАТ

Выпускная квалификационная работа магистра 64 стр., 12 рис., 66 источников, 2 прил.

КРАУЛИНГ ДЕПРЕССИВНОГО КОНТЕКСТА В СОЦИАЛЬНЫХ СЕТЯХ, ПРИМЕНЕНИЕ РЕКУРРЕНТНЫХ СЕТЕЙ

Ключевые слова: депрессия, анализ текста, социальные сети, машинное обучение, нейронные сети, краулинг, выявление аномалий, русскоязычные данные, эвристическая разметка, психоэмоциональное состояние, обработка естественного языка, механизм внимания, attention mechanism

Объект исследования – текстовый контент социальных сетей в контексте выявления психологического состояния пользователей.

Предмет исследования – методы анализа текстовых данных с целью выявления депрессивного контекста, в частности, с применением рекуррентных нейронных сетей.

Цель данной работы – разработка программного обеспечения для краулинга страниц пользователей социальных сетей с целью выявления депрессивного контекста.

Для достижения цели требуется решить следующие задачи:

1) Анализ существующих исследований и решений в области выявления депрессивного контекста в социальных сетях с использованием машинного обучения,

2) Разработка и реализация алгоритмов, использующих нейронные сети для выявления депрессивного контекста,

3) Определение набора данных для обучения модели,

4) Разработка интерфейса для взаимодействия с моделью, в нашем случае краулер.

Выпускная квалификационная работа выполнена в текстовом редакторе и представлена в электронном виде.

СОДЕРЖАНИЕ

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

ML — Machine Learning / Машинное обучение

NLP — Natural Language Processing / Обработка естественного языка

RNN — Recurrent Neural Network / Рекуррентная нейронная сеть

LSTM — Long Short-Term Memory / Долговременная краткосрочная память

BiLSTM — Bidirectional Long Short-Term Memory / Двунаправленная LSTM

CNN — Convolutional Neural Network / Сверточная нейронная сеть

BERT — Bidirectional Encoder Representations from Transformers / Бидирективные представления с использованием трансформеров

SBERT — Sentence-BERT / Модификация BERT для представления предложений

GRU — Gated Recurrent Unit / Ячейка с управляемыми вентилями

TF-IDF — Term Frequency-Inverse Document Frequency / Частота термина — обратная частота документа

SVM — Support Vector Machine / Метод опорных векторов

XGBoost — Extreme Gradient Boosting / Улучшенный градиентный бустинг

POS — Part-of-Speech / Часть речи

LDA — Latent Dirichlet Allocation / Метод латентного размещения Дирихле

API — Application Programming Interface / Программный интерфейс приложения

SMOTE — Synthetic Minority Over-sampling Technique / Метод синтетического увеличения выборки меньшинства

ReLU — Rectified Linear Unit / Прямолинейная единичная функция активации.

Sigmoid — Сигмоидальная функция активации.

JSON — JavaScript Object Notation / Формат обмена данными.

CLS — [CLS] Token в BERT — классификационный токен.

USE — Universal Sentence Encoder / Универсальный кодировщик предложений.

PHQ-9 — Patient Health Questionnaire-9 – Девятибалльная шкала оценки депрессии

OvR (One-vs-Rest) — это стратегия многоклассовой классификации, при которой для каждого класса обучается отдельный бинарный классификатор: он отличает данный класс от всех остальных (то есть «один против всех»).

ВВЕДЕНИЕ

Депрессия — это психическое расстройство, характеризующееся устойчивым снижением настроения, ангедонией, когнитивными нарушениями и соматическими симптомами, которые сохраняются не менее двух недель и нарушают социальное функционирование. Согласно МКБ-10 (принятой в российской клинической практике):

Депрессивный эпизод проявляется снижением настроения, утратой интересов, повышенной утомляемостью, снижением самооценки, идеями виновности, нарушениями сна и аппетита [1]. Основными симптомами являются депрессивное настроение почти ежедневно и большую часть дня, отчетливое снижение интереса или удовольствия от деятельности, которая была ранее приятна, а также снижение энергии и повышенная утомляемость.

Особенно остро проблема депрессивных состояний проявляется среди молодых людей, активно использующих социальные сети для общения, обмена информацией и самовыражения. Депрессия приводит к физическим и эмоциональным проблемам и влияет на работоспособность человека [2].

Определение депрессивного контекста в текстах пользователей социальных сетей представляет собой задачу бинарной классификации, суть которой — отнести текст к одной из двух категорий: содержит ли он признаки депрессии или нет. Подобная постановка задачи широко применяется в сфере анализа текстов с использованием методов машинного обучения.

Алгоритмы машинного обучения — такие как логистическая регрессия, деревья решений, ансамблевые методы и нейросетевые архитектуры — позволяют обучать модели на подобных признаках и впоследствии автоматически определять, относится ли новый текст к категории с депрессивной окраской.

В ходе анализа существующих методов было выявлено, что модели на основе рекуррентных нейронных сетей, особенно BiLSTM с механизмом внимания (Attention), демонстрируют высокую эффективность при анализе

текстовых данных. Поэтому основное внимание в данной работе будет уделено исследованию и применению именно этой архитектуры.

Гипотеза исследования – применение модели BiLSTM с механизмом внимания (Attention) позволяет учитывать лингвистические и психологические особенности русскоязычных текстов, обеспечивая высокую точность выявления депрессивного контекста в социальных сетях.

Актуальность темы – современные социальные сети играют важную роль в жизни общества, предоставляя пользователям платформу для общения, обмена информацией и выражения эмоций. В связи с ростом числа пользователей и объема генерируемого контента возрастает необходимость в автоматическом анализе текстов с целью выявления депрессивного контекста. Депрессивные состояния пользователей могут быть предвестниками серьезных проблем психического здоровья. На данный момент существует множество различных исследований в данной области, однако основная часть сконцентрирована в англоязычном сегменте, и результаты англоязычных исследований не учитывают особенности русскоязычной семантики и лингвокультурных нюансов, что существенно влияет на точность выявления депрессивного контекста в русскоязычных сообщениях. Потенциальными пользователями разработанных в рамках исследования инструментов могут стать кадровые службы коммерческих организаций, для которых психологическое состояние сотрудников имеет критическое значение. Так же данная разработка может послужить как дополнение к более комплексным задачам, таким как разметка для поиска или рекомендательных систем.

Один из ключевых факторов, подчеркивающих уникальность данной работы, заключается в ее фокусе на русскоязычном сегменте, тогда как основная часть исследований в этой области ориентирована на англоязычные корпуса, что приводит к игнорированию лингвистических и культурных особенностей русского языка и снижает точность и воспроизводимость результатов в локальном контексте.

Проблематика – ручной анализ текстов в социальных сетях является трудоемким и субъективным процессом, что делает его малопригодным для масштабного мониторинга эмоционального состояния пользователей. Современные алгоритмы машинного обучения позволяют автоматизировать этот процесс и повысить его эффективность, а так же позволяют осуществить переход от субъективных к более объективным способам диагностики с опорой на реальные поведенческие признаки [3]. Вместе с традиционными методами машинного обучения внедрение нейросетевых методов прогнозирования поможет повысить точность модели, поскольку нейронные сети могут выявлять сложные нелинейные зависимости между признаками и целевой переменной. Это может привести к более точным и эффективным моделям прогнозирования, способным учитывать разнообразные и сложные характеристики пользователей [4].

ОСНОВНАЯ ЧАСТЬ

1. Анализ существующих исследований и решений в области выявления депрессивного контекста в социальных сетях

1.1 Теоретическая часть моделей

Данный раздел направлен на решение первой задачи исследования – анализ существующих исследований и методов выявления депрессивного контекста в текстах социальных сетей.

Современные задачи автоматической классификации текстов — будь то определение тональности, выявление депрессивного контекста или анализ эмоциональной окраски — немыслимы без применения эффективных алгоритмов машинного обучения и нейросетевых архитектур. Самые эффективные и популярные решения сейчас завязаны на использовании таких алгоритмов как LSTM, CNN, BERT, XGBoost и SVM и др., поэтому правильное понимание принципов работы этих моделей играет ключевую роль при выборе подхода для решения конкретной задачи. Как отмечают Pinto и Parente в своем всестороннем обзоре подходов по выявлению депрессии [5], именно эти методы являются наиболее распространёнными в задачах анализа психоэмоционального состояния, включая диагностику депрессии, за счёт их высокой точности и способности обрабатывать как структурированные, так и неструктурированные данные.

Каждая из этих моделей основана на различных архитектурных и математических предпосылках, имеет собственные сильные стороны и ограничения в контексте текстовой классификации. Так, одни модели лучше справляются с последовательной природой текста, другие — с локальными паттернами или структурированными признаками. Различия в их устройстве и применимости напрямую влияют на точность, устойчивость и интерпретируемость конечных результатов.

RNN (Recurrent Neural Network) — это нейронная сеть, которая работает с последовательными данными. Она запоминает информацию о предыдущих

шагах последовательности, что делает её полезной для обработки текстов, где порядок слов важен. Однако стандартные RNN страдают от проблемы затухающего градиента при работе с длинными последовательностями, из-за чего эффективность обучения резко снижается. Как показано в теоретическом и эмпирическом исследовании [6], при увеличении длины последовательности градиенты в стандартной RNN экспоненциально стремятся к нулю, в то время как архитектуры LSTM и GRU демонстрируют значительно лучшую устойчивость благодаря своей внутренней структуре памяти и управляющим элементам [6]. Текст подаётся как последовательность токенов (слов или символов). RNN обрабатывает их по одному, обновляя внутреннее состояние на каждом шаге. Последнее состояние (или объединённые состояния) передаётся на выходной слой классификации (обычно Softmax), который предсказывает метку класса текста.

LSTM (Long Short-Term Memory) — усовершенствованная разновидность RNN, специально разработанная для преодоления проблемы затухающего градиента. Благодаря механизму «затворов», LSTM может эффективно сохранять или забывать информацию, что делает её особенно полезной для анализа текста, речи и временных рядов. В статье [7] сравниваются методы машинного обучения, включая XGBoost и LSTM, для прогнозирования фондового рынка. Результаты показывают, что модель LSTM продемонстрировала наивысшую точность и лучшее соответствие модели, подчёркивая её способность улавливать временные зависимости в данных.

CNN (Convolutional Neural Network) – хотя CNN больше известны в компьютерном зрении, они также применяются в NLP. В текстовых задачах CNN извлекает локальные признаки (например, сочетания слов или фраз) с помощью свёрток. Часто используется в задачах классификации текста, особенно когда важен контекст ближайших слов. Текст представляется как матрица эмбеддингов. Свертки проходят по этой матрице, извлекая локальные

признаки (n-граммы), после чего применяется пулинг (сжатие) и результат передаётся на полносвязный слой классификации. Хорошо работает для коротких и структурированных текстов. CNN и RNN позволяет эффективно захватывать как локальные, так и последовательные особенности текста, что приводит к повышению точности классификации [8].

BERT (Bidirectional Encoder Representations from Transformers) — это мощная трансформер-модель от Google, обученная на больших текстовых корпусах. Главное отличие — двунаправленное обучение: модель понимает контекст слова с обеих сторон. BERT особенно хорошо работает в задачах понимания языка, классификации, поиска и извлечения информации. Это достигается благодаря способности трансформеров учитывать двунаправленные зависимости слов и учитывать более сложные семантические и контекстные взаимосвязи в тексте. Данные модели превосходят традиционные методы, демонстрируя более высокие показатели точности и устойчивости при определении достоверности заголовков и текстовых сообщений [9]. Текст токенизируется и проходит через слои трансформера. Вектор специального токена [CLS] используется как агрегированное представление всего текста и подаётся на классификационный слой, который предсказывает метку (например, депрессивный/не депрессивный) [10]. Этот вектор подаётся на классификационный слой, который предсказывает метку (например, депрессивный/не депрессивный).

XGBoost (Extreme Gradient Boosting) [11] — это бустинговый алгоритм, построенный на деревьях решений. Он эффективен, быстро работает, обладает высокой точностью и часто используется в табличных задачах. В NLP может применяться при наличии извлечённых числовых признаков текста (TF-IDF, частотные характеристики и др.). Перед подачей текст преобразуется в числовые признаки (TF-IDF, частота слов, длина предложений и т.п.). XGBoost обучается на этих признаках, строя ансамбль деревьев, которые принимают

решение на основе значений признаков. Подходит для табличного представления текстов.

SVM (Support Vector Machine) — это алгоритм, который ищет гиперплоскость, наилучшим образом разделяющую классы в признаковом пространстве. Он хорошо работает в задачах бинарной классификации, особенно на разреженных текстовых признаках. Часто применяется с TF-IDF или мешком слов. Текст также сначала преобразуется в векторное представление (например, TF-IDF). SVM ищет гиперплоскость, которая разделяет классы в этом пространстве. Используется для бинарной или многоклассовой классификации и особенно эффективен при работе с разреженными данными. В исследовании [12] подчеркивается эффективность SVM в задачах классификации текста.

1.2 Анализ существующих исследований

Анализ социальных медиа предоставляет уникальные возможности для выявления ранних признаков депрессии у пользователей. Различные подходы, такие как машинное обучение, глубокие нейронные сети и традиционные статистические методы, используются для автоматического анализа текстов, изображений, метаданных и поведенческих характеристик. Данный обзор направлен на анализ существующих решений, выделение их преимуществ и недостатков, а также обсуждение направлений будущих исследований в этой области.

1.2.1 Классические методы машинного обучения

Некоторые исследования используют традиционные алгоритмы машинного обучения. Классические методы машинного обучения представляют собой алгоритмы, основанные на анализе признаков. Они включают логистическую регрессию, опорные вектора (SVM), случайные леса (Random Forest) и метод наивного Байеса. Эти модели требуют тщательной предварительной обработки данных и извлечения информативных признаков,

но обладают высокой интерпретируемостью по сравнению с глубоким обучением.

В работе с использованием методов на основе текстовых и поведенческих признаков [13] анализируется активность пользователей, частота публикаций, эмоциональный тон сообщений, текстовые и поведенческие признаки пользователей. Используются такие алгоритмы, как логистическая регрессия, SVM и случайный лес.

В исследовании Obagbuwa и соавт. [14] была проведена сравнительная оценка четырёх классических моделей машинного обучения (Logistic Regression, SVM, XGBoost, Random Forest) на задаче анализа депрессивных настроений в твитах. Авторы использовали объединённые датасеты с платформы Kaggle, содержащие размеченные по полярности сообщения. Внимание уделено тщательной предобработке текста: лемматизация, очистка от ссылок, повторов, стоп-слов и создание биграмм. По результатам эксперимента, наивысшую точность (96.3%) при минимальном времени исполнения (0.29 сек) показала логистическая регрессия, что подчёркивает её применимость в задачах реального времени. Авторы делают акцент на том, что корректная работа с данными на этапе препроцессинга способна компенсировать простоту модели, обеспечивая конкурентное качество классификации.

Систематические обзоры [15, 16] обобщают существующие методы, выявляя трудности, связанные с интерпретацией и разметкой данных. Авторы [15] анализируют 51 исследование, отмечая, что выбор алгоритма, как правило, зависит от доступности и качества данных. В [16] приводится обобщение различных техник извлечения признаков и моделей, используемых в задачах выявления депрессии в социальных медиа.

Ding и соавт. [17] предложили метод выявления депрессии у студентов колледжа на основе анализа текстов в китайской социальной сети Sina Weibo. В исследовании используется комбинация глубокой нейросети для

автоматического извлечения признаков и метода интеграции опорных векторов (DISVM), построенного с применением AdaBoost. Среди признаков учитываются частотность слов, типичные эмодзи, поведенческие характеристики (активность, число подписчиков и т.д.). Проведённый эксперимент показал, что DISVM превосходит по точности традиционные модели (SVM, KNN, RBF-NN), достигая точности до 90.1%. Отдельно отмечается важность выбора временного диапазона: наилучшие результаты достигаются при анализе поведения за 24 месяца. Работа демонстрирует практическую применимость классических методов ML к задачам ментального здоровья в онлайн-среде.

В работе [18] в дополнении к использованию классических методов машинного обучения, N-грамм и TF-IDF, применяется метод тематического моделирования позволяющий выделять темы, связанные с депрессивным состоянием, из текстов в социальных сетях - Latent Dirichlet Allocation (LDA).

LDA - это пробабилистическая (статистическая) модель тем, которая используется для выделения скрытых тем в больших коллекциях текстов (корпусах). Проще говоря LDA, пытается автоматически понять, о чём тексты, разбивая их на темы без необходимости ручной разметки.

Дополнительные сложности при применении классических методов машинного обучения возникают при анализе текстов на русском языке. Согласно исследованию [19], высокая степень морфологической изменчивости и синтаксической гибкости русского языка затрудняет построение стабильных признаковых пространств. Авторы подчёркивают, что в условиях ограниченности размеченных корпусов и невозможности использовать крупные предобученные модели, методы на основе TF-IDF, n-грамм и POS-тегов остаются наиболее эффективными и воспроизводимыми в задачах анализа тональности и эмоционального контекста русскоязычных текстов.

Несмотря на свою интерпретируемость, классические алгоритмы уступают нейросетям в умении учитывать контекст и семантику текста. Также они зависят от качества инженерии признаков, что делает их менее гибкими.

1.2.2 Применение гибридных методов

Гибридные методы сочетают в себе возможности различных моделей для повышения точности детекции депрессии. Это может быть комбинация глубокого обучения с традиционными методами машинного обучения, а также использование мультимодальных данных (например, текст + поведенческие характеристики пользователя). Такие подходы позволяют учитывать не только текстовые признаки, но и активность пользователя, временные паттерны и метаданные публикаций. Сочетание различных подходов дает лучшие результаты.

В работе [20] использование трехуровневой архитектуры BiLSTM для суммаризации показывает внушительные результат. Архитектура позволяет учитывать сложные зависимости и обеспечивает устойчивые результаты.

Совмещение классификации и объяснения на основе клинических шкал, таких как PHQ-9, обеспечивает как точность, так и прозрачность в выявлении депрессии [21].

Также стоит отметить актуальность применения предварительно обученных глубоких сетей с использованием Transfer Learning в работе [22], где рассматривалась детекция депрессии. Там же рассматривались и мультимодальные подходы, из них выделяются Fusion-модели, которые комбинируют текстовые признаки с метаданными (например, частотой активности пользователя, временем публикаций) для более точного предсказания депрессии.

В работе [23] представлен обширный обзор современных методов ИИ для выявления депрессии на основе анализа социальных сетей. Особое внимание уделено мультиклассовой классификации эмоционального

состояния пользователей с использованием как текстовых данных, так и эмодзи/эмотиконов. В ходе обзора 101 публикации установлено, что гибридные архитектуры — например, комбинации CNN + BiLSTM, LSTM + Attention, FastText + BiGRU — демонстрируют наивысшую точность (до 93%) при выявлении депрессивного контекста. Отдельно подчеркивается роль мультимодальных признаков и эмодзи как значимых индикаторов эмоционального состояния. Авторы делают вывод о перспективности комплексных моделей, сочетающих глубокое обучение с лингвистическим анализом и расширенной векторизацией, в задачах анализа психоэмоционального состояния пользователей.

В рамках исследования [24] была предложена модель DepressionNet, сочетающая в себе поведенческие и текстовые признаки пользователя для задачи выявления депрессии в Twitter. В основе модели лежит оригинальный подход, включающий гибридную автоматическую суммаризацию пользовательских твитов (экстрактивная на основе BERT и абстрактивная через DistilBART) с последующей обработкой текстов в CNN-BiGRU-архитектуре с вниманием. Дополнительно используется отдельная ветвь, анализирующая поведение пользователя: частоту публикаций, соцсвязи, эмоциональные характеристики и тематическую структуру (LDA). Слияние признаков на уровне высокоуровневого представления (late fusion) позволило достичь F1-метрики 0.912 и точности 90.1%, превзойдя существующие модели. Работа подчёркивает эффективность совместного анализа контекста сообщений и цифрового поведения пользователей для раннего выявления депрессии.

1.2.3 Лингвистический и психологический анализ

В данном разделе рассматриваются исследования, в которых анализ депрессии осуществляется на основе интерпретируемых лингвистических и психологических признаков, независимо от используемой архитектуры

модели. Такие признаки, как частота использования местоимений первого лица, эмоциональная окраска текста, синтаксические особенности и тематическая структура сообщений, являются индикаторами психоэмоционального состояния пользователя и активно используются в задачах классификации депрессивного контекста.

В исследовании [25] анализировались посты пользователей социальной сети Reddit. Авторы выявляли депрессивные паттерны на основе частотного распределения ключевых слов и тематического моделирования. Модель классификации основывалась на признаках, таких как использование негативно окрашенной лексики, частота глаголов, выражающих безнадежность, и наличие лексем, указывающих на изоляцию. Однако в работе не приводится информация о применяемом объеме данных или методах балансировки классов.

В статье [26] особое внимание уделено выявлению признаков психических расстройств на основе анализа текстов из Twitter. Использовались предварительно обученные языковые модели и методы transfer learning, однако ключевым моментом исследования являлся именно лексико-психологический анализ. В выборке были выделены наиболее значимые маркеры депрессии: преобладание местоимений «я», частое употребление слов с негативной коннотацией и сниженная вариативность лексики. Авторы проводили классификацию на основе этих признаков, не ограничиваясь архитектурными особенностями моделей.

Особое внимание к лингвистическим аспектам и методологическим рискам анализа психического состояния в социальных сетях уделяется в обзоре Cao и соавт. [27], где рассмотрено 47 исследований, использующих ML и DL для выявления признаков депрессии. Авторы показывают, что в 77% работ не обрабатываются отрицания, несмотря на их критическую роль в анализе настроений. Также выявлены другие лингвистические и социокультурные перекосы — почти все исследования опираются на

англоязычные тексты, преимущественно из Twitter, что ограничивает обобщаемость. Психолингвистические особенности (ирония, эмоциональная экспрессия, грамматические конструкции) часто игнорируются при построении признаков. Обзор подчёркивает, что глубокие модели не могут компенсировать методологические упрощения в препроцессинге и аннотации, и делает вывод о необходимости учитывать когнитивную сложность языка при построении интерпретируемых систем анализа ментального состояния.

Также в ряде обзоров [15, 16] подчёркивается значимость применения психологически обоснованных признаков для интерпретируемой классификации. Авторы подчёркивают, что такие признаки, как частота публикаций, использование определённых словоформ, длина предложений, количество знаков препинания и эмодзи, могут быть полезны как маркеры эмоционального состояния.

Таким образом, лингвистические и психологические признаки выступают важным направлением в задаче автоматического анализа депрессии, поскольку обеспечивают интерпретируемость и могут использоваться как в классических алгоритмах, так и в составе нейросетевых решений. Включение таких признаков повышает надёжность интерпретации результатов и способствует более точной верификации классификационных моделей.

1.2.4 Подходы на основе глубокого обучения

Глубокое обучение стало ключевым направлением в задачах автоматического выявления признаков депрессии в текстах социальных сетей. Модели, основанные на нейронных сетях, способны автоматически извлекать сложные иерархические представления из текста, учитывая как семантические, так и синтаксические особенности. Это позволяет более точно моделировать скрытые паттерны, характерные для депрессивного лингвистического поведения.

В исследовании [28] предложили модель для распознавания текстовых следствий, основанную на двунаправленной LSTM (BiLSTM) с механизмом внутреннего внимания (Inner-Attention). Модель обучалась на корпусе SNLI, содержащем 570 тысяч пар предложений. Особенностью подхода является использование внутреннего внимания для улучшения представления предложений, что позволило достичь высокой точности классификации. В статье не указаны методы балансировки классов или специфические особенности сбора данных. Кузнецов Р. С. в своей работе [29] применил модель BiLSTM с механизмом внимания для прогнозирования биржевых котировок Amazon. Модель обучалась на временных рядах цен акций, используя исторические данные. Особенностью подхода является сочетание BiLSTM и механизма внимания для захвата временных зависимостей в данных. Однако в статье не указаны методы балансировки классов или специфические особенности сбора данных.

В работе [30] разработали гибридную модель SBERT-CNN для выявления пользователей Reddit с признаками депрессии. Модель использует эмбеддинги предложений из SBERT и извлекает локальные признаки с помощью сверточной нейронной сети. Особенностью подхода является сочетание семантического представления текста и извлечения локальных признаков. Однако в статье не указаны объём данных, методы балансировки классов и способы получения данных для обучения.

Kour и Gupta [31] предложили гибридную модель, объединяющую сверточные нейронные сети (CNN) и двунаправленные LSTM (BiLSTM) для предсказания депрессии на основе твитов пользователей. Модель обучалась на выборке твитов, содержащей как депрессивные, так и недепрессивные сообщения. Особенностью подхода является использование CNN для извлечения локальных признаков и BiLSTM для захвата долгосрочных зависимостей в тексте. В данной статье нет описания примененных методов балансировки данных и их происхождения.

Исследование [32] сравнивает модели LSTM и BERT на небольшом корпусе текстов, собранном для задач классификации намерений в чат-ботах. Результаты показали, что LSTM значительно превосходит BERT по точности на малых выборках и требует меньше времени на обучение. В статье не указаны точный объём данных, методы балансировки классов и способы получения данных для обучения.

Таким образом, подходы на основе глубокого обучения демонстрируют высокую эффективность в задачах детекции депрессии, особенно при наличии больших объемов данных и использовании предобученных моделей. Для достижения интерпретируемости и учета психологических аспектов рекомендуется комбинировать эти подходы с лингвистическим и психологическим анализом.

Однако подобные модели требуют значительного объема размеченных данных и вычислительных ресурсов, что ограничивает их применение в условиях ограниченных ресурсов или на малых выборках. Кроме того, интерпретируемость таких моделей часто вызывает затруднения [33, С. 215].

В исследовании [34] представлены методы выявления депрессии в блогах на тайском языке с использованием современных нейросетевых архитектур, включая модели с механизмом внимания. На корпусе из более чем 33 тыс. размеченных сообщений авторы сравнили эффективность различных подходов: от классических моделей (SVM, Наивный Байесовский классификатор) и традиционных нейросетей (CNN, LSTM, Bi-GRU) до трансформеров (M-BERT и XLM-RoBERTa). Наивысшие результаты достигнуты с помощью XLM-RoBERTa (точность 79.12%, F1-метрика 0.8016), что авторы объясняют эффективным мультилингвальным обучением и способностью модели извлекать сложные контекстуальные зависимости в условиях ограниченности ресурсов. Данное исследование подчёркивает потенциал использования трансформеров и attention-механизмов в задачах

анализа депрессивного контекста, особенно для языков с ограниченным набором данных для обучения.

В работе [35] предложена инновационная многоуровневая архитектура на основе BERT и BiGRU с механизмом внимания, ориентированная на детекцию депрессии в текстах из социальных сетей. В отличие от большинства подходов, авторы дополнили модель признаками, связанными с аффективной сферой (эмоции), а также социально-психологическими маркерами — бранной лексикой и моральными категориями, извлекаемыми на основе LIWC и Moral Foundations Dictionary. Эти признаки внедряются в модель через механизм позднего объединения (late fusion), усиливая семантические представления BERT. Эксперименты на датасетах RSDD и Pirina показали, что добавление эмоциональных и нормативных признаков значительно повышает точность классификации (до +6.73% F1), демонстрируя важность анализа социального контекста в задачах ментального здоровья. Подход авторов выделяется своей интеграцией поведенческих и этических сигналов в рамки нейросетевой архитектуры, что открывает новые направления в интерпретируемом ИИ для психодиагностики.

1.2.5 Локальные исследования на русскоязычных данных

Анализ русскоязычных социальных сетей остается актуальной задачей. Русскоязычные социальные сети, такие как ВКонтакте и Telegram, представляют собой отдельное исследовательское направление. Из-за языковых особенностей и меньшего количества доступных размеченных данных анализ депрессивного контекста в русскоязычном сегменте требует адаптации существующих моделей. В этом разделе рассматриваются работы, посвященные использованию RuBERT, DeepPavlov и других специализированных моделей для анализа русскоязычного контента.

Так в исследовании [36] производится выявление депрессии на основе анализа постов и комментариев пользователей ВКонтакте. Применение

методов машинного обучения для детекции депрессии в русскоязычном сегменте [4, 37] исследуют специфики языка и особенностей русскоязычного интернет-пространства.

Одним из наиболее масштабных и репрезентативных локальных исследований является работа Сметанина и Комарова [38], в которой были протестированы модели M-BERT, RuBERT и Multilingual USE на семи открытых русскоязычных корпусах, включая RuSentiment, LINIS Crowd и RuTweetCorp. Авторы не только достигли новых state-of-the-art результатов на нескольких задачах, но и провели подробный анализ трудностей, возникающих при обработке русскоязычных текстов: от несбалансированной разметки до сложности с интерпретацией нейтральных классов и особенностей токенизации эмодзи.

В рамках локального контекста важно учитывать специфику обработки русскоязычных текстов. В исследовании Чернышова [39] рассматривается метод синтаксического анализа пользовательских запросов с помощью BiLSTM-модели, обученной на русскоязычном корпусе (UD_Russian-SynTagRus). В модели используется механизм парсинга зависимостей, позволяющий строить иерархическую структуру запроса с сохранением смысловых связей. Хотя сама работа ориентирована на задачи семантического поиска, предложенный подход может быть адаптирован для анализа поведенческих и эмоциональных паттернов в пользовательских текстах. Это особенно актуально в контексте диагностики депрессивных состояний по русскоязычным данным, где важно учитывать морфологическую сложность языка и семантическую насыщенность.

1.3 Метрики

При рассмотрении различных подходов, будет вестись сравнение результатов в распространённых метриках, использованных в исследованиях, а именно F1 и Accuracy:

Accuracy (Точность классификации) – показывает долю правильно классифицированных примеров среди всех примеров. Вычисляется по формуле:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

Где TP (True Positive) – количество верно предсказанных положительных примеров,

- TN (True Negative) – количество верно предсказанных отрицательных примеров,

- FP (False Positive) – количество ложно положительных предсказаний,

- FN (False Negative) – количество ложно отрицательных предсказаний.

Однако Accuracy не всегда является надежной метрикой, особенно при несбалансированных данных.

F1-score – учитывает как полноту, так и точность, что делает её более надежной в задачах с несбалансированными классами.

Precision показывает, насколько точно модель определяет положительные классы, то есть, какая доля предсказанных моделью положительных примеров (депрессивных случаев) действительно является таковой.

$$Precision = \frac{TP}{TP + FP}, \quad (2)$$

Recall показывает, насколько хорошо модель находит все реальные положительные примеры, т.е. какую долю всех депрессивных случаев модель успешно предсказала.

$$Recall = 2 \times \frac{TP}{TP + FN}, \quad (3)$$

А сам F1-score вычисляется по формуле:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \quad (4)$$

F1-score особенно полезен, когда важно сбалансировать Precision и Recall, например, при выявлении депрессивного контекста, где ошибки в предсказании могут иметь серьезные последствия.

1.4 Результаты анализа

В результате проведённого анализа были детально рассмотрены и проанализированы различные подходы к решению задачи автоматического выявления депрессивного контекста в текстах социальных медиа. Исследование охватывает современные нейросетевые архитектуры (RNN, LSTM, CNN, BERT, SBERT), классические алгоритмы машинного обучения (SVM, логистическая регрессия, случайные леса), лингвистические и психологические методы, а также гибридные и мультимодальные подходы.

Далее в работе представлены результаты эмпирического анализа эффективности рассмотренных моделей в виде таблицы с количественными показателями их работы и сравнительный анализ, позволяющий выбрать оптимальный подход для выявления депрессивного контекста в текстах социальных медиа.

Данные результаты включают в себя исследование, модель, метрики F1 и Accuracy, и датасет на котором проводились эксперименты.

Стоит отметить, что несмотря на демонстрируемые высокие показатели точности, большинство моделей в таблице тестировались на англоязычных данных и не всегда устойчивы к смещению в выборке. Также стоит учитывать, что лишь немногие исследования уделяют внимание реальной воспроизводимости результатов или оценке на внешних валидационных выборках.

С подготовленной таблицей можно ознакомится в приложении А.

Таблица показывает, что классические методы машинного обучения могут быть эффективны, однако они уступают нейросетям в точности классификации. Методы анализа текста, такие как BERT и BiLSTM,

демонстрируют более высокие результаты. Логистическая регрессия и SVM дают удовлетворительные результаты, но их производительность может варьироваться в зависимости от качества входных данных и используемых признаков.

В рассмотренных работах используется сравнительно небольшой объем данных. При этом датасеты часто страдают от несбалансированности классов — депрессивные тексты чаще всего составляют менее 30% общей выборки, что требует использования специальных техник (взвешивание классов, oversampling). В ряде работ балансировка не описана, что может снижать обобщающую способность моделей.

Большинство современных исследований и моделей ориентировано на англоязычный сегмент социальных сетей (Twitter, Reddit, Facebook). Такие модели, как BERT и SBERT, показывают высокие результаты на крупных англоязычных корпусах, однако имеют ограниченную применимость к русскоязычным данным [4].

Методы машинного обучения находят широкое применение в поведенческой диагностике депрессии, однако исследования сталкиваются с проблемами воспроизводимости, качества данных и этических аспектов [40], что подчеркивает важность исследований различных методов, на различного рода данных.

В совокупности результаты анализа демонстрируют, что наибольшую эффективность в задаче автоматической детекции депрессивного контекста в текстах социальных сетей показывают архитектуры, способные учитывать последовательную природу данных и сложные контекстуальные зависимости. К числу таких решений относятся рекуррентные нейронные сети, в частности модификации на основе LSTM и BiLSTM, а также трансформерные модели, представленные архитектурами BERT и SBERT. Указанные подходы различаются по вычислительной сложности, требованиям к объёму обучающих данных, а также степени интерпретируемости получаемых

результатов. Учитывая выявленные особенности и эмпирические показатели эффективности, в последующих разделах будет осуществлён сравнительный анализ данных классов моделей с целью обоснованного выбора архитектуры, наиболее релевантной задачам выявления депрессивного контекста в русскоязычном сегменте социальных сетей.

2. Выбор модели определение метрик для выявления депрессивного контекста, данные для обучения модели

2.1 Выбор модели

В данном разделе решается вторая задача исследования – выбор и разработка модели на основе нейронных сетей, включая настройку гиперпараметров и выбор архитектуры.

Задача выявления депрессивного контекста в пользовательских текстах социальных сетей требует применения таких архитектур, которые эффективно работают с естественным языком и способны учитывать сложную структуру и контекстуальные зависимости. В современном научном сообществе наибольшее распространение получили два направления: рекуррентные нейросетевые архитектуры, включая LSTM и BiLSTM, а также трансформерные модели, такие как BERT и SBERT. Каждое из этих направлений имеет как преимущества, так и ограничения, что требует их подробного сопоставительного анализа.

Модели на основе рекуррентных нейронных сетей (RNN) обладают способностью учитывать последовательную природу текста, так как каждое следующее состояние сети зависит от предыдущего. Однако стандартные RNN имеют существенные ограничения, связанные с проблемой затухающего градиента, из-за чего теряется информация о первых элементах длинных последовательностей. Для преодоления этого недостатка были разработаны архитектуры LSTM (Long Short-Term Memory), обладающие механизмом «долгосрочной памяти», позволяющим хранить информацию на больших интервалах и избирательно запоминать или забывать входные сигналы.

На концептуальном уровне различия между стандартной рекуррентной сетью и LSTM хорошо иллюстрируются на рисунке 1. Слева изображена архитектура классической RNN, где данные проходят через «рабочую память» без механизма долговременного хранения, что делает модель уязвимой к потере контекста на больших текстах. Справа представлена схема LSTM, в

которой, помимо рабочей памяти, задействован отдельный путь долгосрочной памяти (long-term memory), позволяющий сохранять важные семантические зависимости между токенами на протяжении всей последовательности.

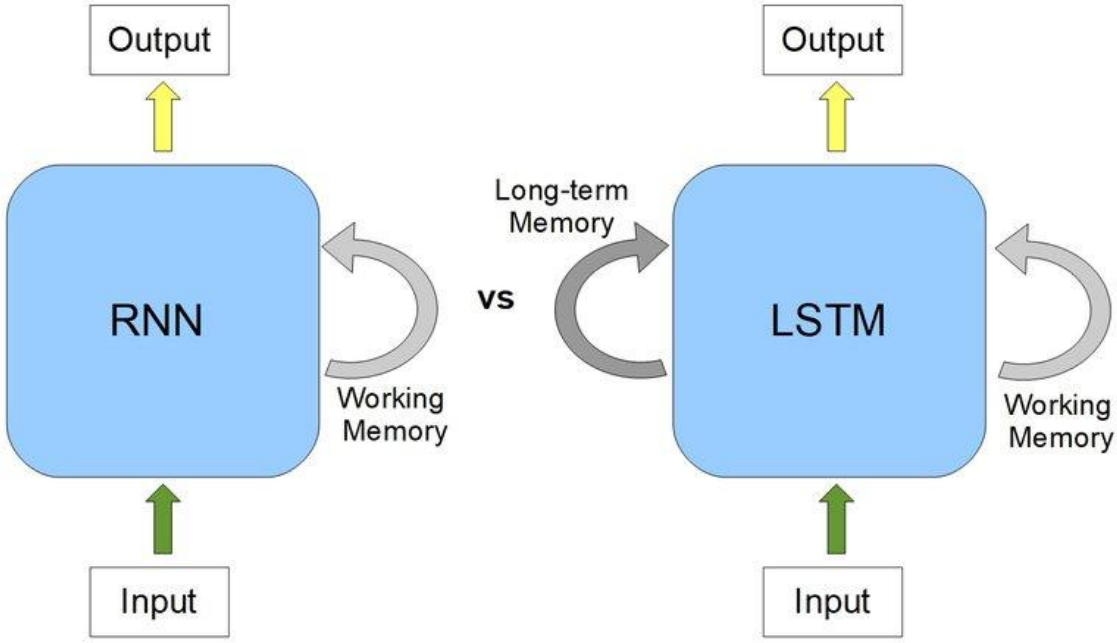

Рисунок 1 - RNN v/s LSTM.[41]. Лицензия: CC BY-ND 4.0. Источник: [42].

Разновидность LSTM — двунаправленная модель BiLSTM — расширяет возможности обработки текста, анализируя его в обоих направлениях, тем самым обеспечивая лучшее улавливание контекста. Практическое применение таких моделей показало высокую эффективность в задачах классификации, включая эмоциональный анализ и предсказание депрессивных состояний, что подтверждается, например, результатами исследования [28], где архитектура BiLSTM с механизмом внимания (attention) обеспечила точность классификации на уровне 84% при работе с данными Twitter.

Сравнительный анализ показывает, что BiLSTM в сочетании с attention-механизмом демонстрирует оптимальный баланс между качеством, устойчивостью и ресурсными затратами. Эта архитектура позволяет не только учитывать контекст с обеих сторон текста, но и направлять внимание модели

на наиболее значимые фрагменты предложения, что особенно важно в задачах психолингвистического анализа. Подобная модель была реализована и протестирована в ряде работ, включая исследование [20], в котором трёхуровневая BiLSTM-сеть достигла F1-метрики 84% при работе с эмоционально окрашенными текстами пользователей социальных сетей.

Важным аспектом выбора модели является ограниченность объема размеченных русскоязычных данных. Несмотря на существование моделей типа RuBERT и Multilingual BERT, их применение на небольших корпусах сопровождается риском переобучения и недостаточной устойчивостью к шумам. Как указывают Сметанин и Комаров [38], при применении RuBERT и Multilingual USE к русскоязычным корпусам выявлены сложности, связанные с токенизацией, интерпретацией нейтральных классов и дисбалансом данных.

Учитывая вышесказанное, а также основываясь на эмпирических результатах, полученных в ряде исследований, выбор модели BiLSTM + Attention представляется обоснованным. Эта архитектура обладает высокой устойчивостью при обучении на малых выборках, сохраняет интерпретируемость решений за счёт механизма внимания и демонстрирует высокие метрики качества при разумных вычислительных затратах. В условиях ограниченных ресурсов, специфики русского языка и необходимости анализа текстов, содержащих признаки депрессии, модель BiLSTM + Attention является наиболее рациональным выбором.

2.2 Данные для обучения моделей

Разработка модели для автоматического выявления депрессивного контекста требует качественных и репрезентативных данных. В англоязычном сегменте интернета доступны открытые датасеты, содержащие тексты с эмоциональной окраской, включая депрессивные посты и комментарии. Так к примеру использование больших языковых моделей в сочетании с

полусупервайзным обучением позволило создать масштабный аннотированный англоязычный датасет и существенно повысить точность систем обнаружения симптомов депрессии на базе социальных медиа [43].

Еще одним примером полного цикла создания датасета с уровнями депрессии является работа Kayalvizhi и Thenmozhi [7], в которой тексты с Reddit были аннотированы двумя экспертами по трёхуровневой шкале: «не депрессивный», «умеренно депрессивный» и «тяжело депрессивный». Корпус составил 16 613 размеченных записей. Авторы провели расчёт коэффициента согласия (Cohen’s Kappa = 0.686), что подтверждает надёжность разметки. Дальнейшая обработка включала векторизацию и обучение моделей классификации, а также балансировку классов с помощью SMOTE. Итоговая модель на Word2Vec + Random Forest достигла F1-метрики 0.877, продемонстрировав применимость классических методов при наличии качественной разметки.

Однако в русскоязычном сегменте (RuNet) подобных публичных датасетов практически нет, что значительно усложняет задачу сбора данных для обучения моделей. В связи с этим в рамках данной работы был разработан парсер для автоматизированного сбора текстов из социальной сети «ВКонтакте» (VK). Данный выбор обусловлен несколькими факторами:

– Отсутствие открытых русскоязычных датасетов. В отличие от англоязычных источников (например, Reddit или Twitter), в RuNet нет широкодоступных размеченных корпусов с депрессивным контентом,

– Доступность постов. Личные сообщения пользователей в соцсетях являются закрытыми и недоступными для анализа. В то же время публичные посты в сообществах и на личных страницах открыты для сбора, что делает их основным источником данных,

– Социальные сети как основная площадка самовыражения. ВКонтакте является одной из крупнейших соцсетей в русскоязычном

интернете, где пользователи активно публикуют личные мысли, переживания и эмоциональные состояния.

Принцип работы парсера: алгоритм начинает работу с указанного пользователя, находит его друзей, фильтрует только открытые профили и загружает их посты. Затем он рекурсивно переходит к друзьям этих друзей, расширяя сеть до заданной глубины.

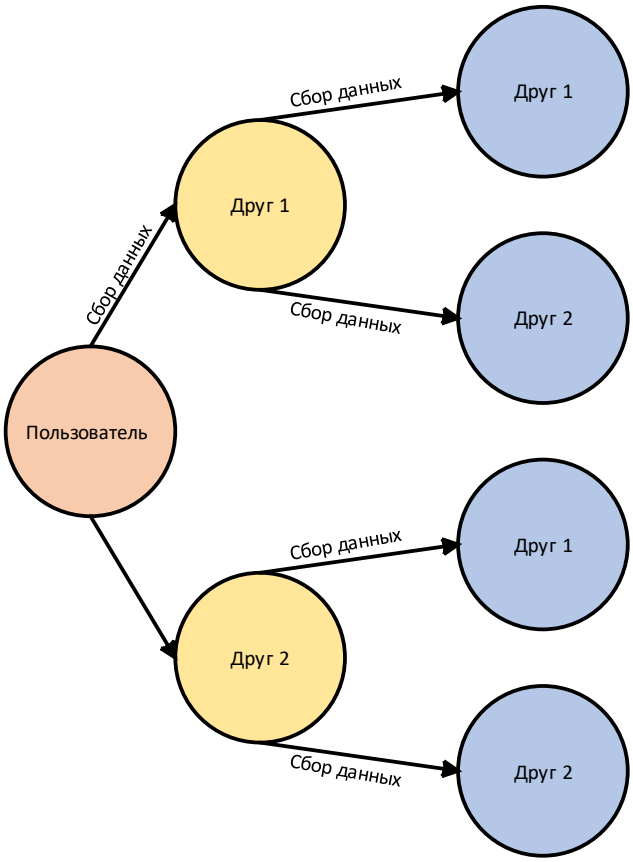

Рисунок 2 – Схема работы парсера.

Все собранные данные (информация о пользователях и их посты) сохраняются в файлы партиями по 1000 объектов. Это помогает избежать потери данных и перегрузки API.

Для обеспечения репрезентативности выборки и включения достаточного числа пользователей с признаками депрессивного поведения был также применён метод отбора участников из специализированных сообществ социальной сети, ориентированных на депрессивную тематику.

Отбор сообществ осуществлялся на основе совокупности критериев, включающих:

– семантический анализ названия и описания (наличие в формулировках ключевых слов, связанных с грустью, депрессией, одиночеством и эмоциональной нестабильностью),

– анализ публикуемого контента (присутствие в постах признаков депрессивных высказываний, цитат, вырезок из литературы и визуального контента в соответствующей стилистике, в том числе мемов),

– характер активности аудитории (наличие пользовательских публикаций, соответствующих тематике сообщества, что подтверждает вовлечённость и релевантность группы).

Таким образом удалось суммарно собрать 140тыс. пользователей с постами.

Для того что бы использовать VK API [44], требуется быть зарегистрированным пользователем и создать “no-code приложение”, процесс создания подробно описан в документации. После создания нужно сгенерировать ключ, с которым уже можно использовать различные методы VK API.

Итоговая реализация парсера была оформлена в виде программной библиотеки и опубликована в репозитории пакетов PyPI (Python Package Index) [45], что обеспечивает её повторное использование и интеграцию в сторонние проекты посредством менеджера зависимостей pip. С реализацией библиотеки можно ознакомится по ссылке [46].

В рамках данного исследования внимание так же уделяется вопросам этики, связанными с обработкой пользовательского контента из социальных сетей. Все данные собираются исключительно из открытых источников, доступных и не защищённых приватными настройками. Сбор информации осуществляется в соответствии с публичной политикой API социальной сети

«ВКонтакте» [47], что позволяет избегать нарушения условий использования платформы.

Несмотря на публичность информации в социальных сетях, её использование в исследованиях может нарушать базовые принципы добровольного участия и информированного согласия. Даже открытые данные требуют деликатного обращения, особенно при наличии риска деанонимизации пользователей [48], поэтому дополнительно, все собранные данные обезличиваются — информация, позволяющая однозначно идентифицировать пользователя (ФИО, ID, ссылки на профили), удаляется до этапа подачи их в модель на обучение. Работа фокусируется исключительно на текстовом содержимом публикаций, которое рассматривается в агрегированном виде.

Также следует отметить наличие российских инициатив в сфере регулирования этичного использования данных. В частности, в 2019 году был принят «Кодекс этики использования данных» [49], разработанный Ассоциацией больших данных совместно с ИРИ и подписанный рядом ведущих компаний. Он закрепляет ключевые принципы — соблюдение прав субъектов данных, прозрачность, минимизацию и недопущение дискриминации.

Исследование не направлено на диагностику или вмешательство в личную жизнь пользователей, а исключительно на развитие методов автоматического анализа текстов для выявления маркеров депрессивного контекста в целях научного моделирования. Такой подход соответствует международной практике, согласно которой анализ открытых данных допустим при соблюдении условий анонимности и научной цели обработки данных [50].

Таким образом, соблюдены базовые требования по этическому использованию пользовательского контента, изложенные в ряде международных работ [51, 52].

Так как в процессе парсинга пользовательского контента с платформы «ВКонтакте» отсутствует заранее заданная разметка депрессивных сообщений, из-за чего возникает необходимость разработать собственную процедуру аннотирования данных. В отличие от англоязычных исследований, использующих self-reported или клинически верифицированные источники — в том числе данные из специализированных сабреддитов Reddit [28, 30] — в русскоязычном сегменте невозможно напрямую определить эмоциональное состояние пользователя на основе метаинформации или принадлежности к сообществу.

Для создания обучающей выборки применим эвристический подход к разметке, основанный на лингвистических и психологических маркерах, описанных в исследованиях [25, 26, 21, 38]. В частности, депрессивным считается текст, содержащий одну или несколько из следующих категорий признаков:

– лексические индикаторы, такие как наличие слов с негативной семантикой («пустота», «бессмысленно», «устал», «не вижу смысла», «никому не нужен», «ненавижу себя» и др.),

– темы изоляции, утраты и эмоционального отдаления, включая упоминания одиночества, разрыва социальных связей или потери мотивации,

– преобладание местоимений первого лица, особенно в сочетании с отрицательной оценочной лексикой, что подтверждено в лингвистических и психолингвистических исследованиях [21, 26],

– структурные особенности текста, такие как низкая вариативность, короткие и однотипные предложения, отсутствие связности,

– упоминания поведенческих и соматических маркеров, включая бессонницу, апатию, тревожность, нарушение пищевого поведения, усталость и другие симптомы, зафиксированные в психиатрической и психологической литературе.

Для автоматизации первичной фильтрации был сформирован набор ключевых слов и фраз, основанный на англоязычных корпусах депрессивных текстов и адаптированный к русскому языку посредством перевода и ручной коррекции, с которым можно ознакомится в источнике [53].

Разметка осуществляется в два этапа. На первом этапе применяется автоматическая эвристическая фильтрация, реализованная через словарь и набор шаблонов. На втором этапе производится частичная ручная верификация случайной подвыборки с целью откалибровать критерии и оценить качество аннотирования. При необходимости возможен переход к формированию вероятностных меток (soft labels) — например, на основе суммарной оценки количества и силы признаков. Однако в данной работе используется бинарная схема: «депрессивный» / «нейтральный».

Применение подобного подхода позволяет, с одной стороны, обеспечить масштабируемость сбора и подготовки данных, а с другой — достичь приемлемой степени достоверности и репрезентативности выборки, пригодной для обучения нейросетевой модели. Практическая применимость подобной схемы аннотирования подтверждена в ряде работ по автоматической диагностике депрессии, включая [28, 26, 21], где аналогичные признаки служили основой для слабо наблюдаемого обучения в задачах психоэмоционального анализа.

В рамках подготовки данных была реализована процедура очистки и структурирования информации о пользователях: очистка текстов пользовательских публикаций с удалением URL-ссылок, специальных символов и лишних пробелов, с приведением текста к нижнему регистру; нормализованы пользовательские профили, включающую извлечение базовых социально-демографических признаков (пол, город, число подписчиков, аттрибуты из раздела «Личное»), статус пользователя и посты, прошедшие очистку. Также проводится агрегация метаданных к постам (дата, количество лайков, комментариев, репостов и просмотров); сформирован итоговый

корпус, состоящего только из тех пользователей, у которых отсутствуют пустые публикации.

Таким образом в отличие от существующих работ, где сбор данных зачастую ограничивается общедоступными корпусами или англоязычными датасетами, в данном исследовании была разработана собственная система парсинга для автоматического сбора данных из социальной сети ВКонтакте, а также специально разработанная методика разметки данных на основе лингвистических и психологических маркеров для обеспечения высокого качества данных и минимизации рисков ошибок. Это позволило собрать уникальный корпус из 85 914 пользователей и 206 805 постов.

С примером итоговых данных можно ознакомится в приложении Б.

2.3 Краулинг

Краулинг (от англ. web crawling) — это автоматизированный процесс обхода веб-страниц с целью сбора информации. Программы, осуществляющие этот процесс, называются веб-краулерами (web crawlers, spiders). Они переходят по ссылкам, загружают содержимое страниц и извлекают данные, которые могут быть использованы для анализа, индексации или хранения [54].

Веб-краулеры широко применяются в различных прикладных и исследовательских задачах. Поисковые системы (например, Google, Bing, Яндекс) используют краулеры для индексирования веб-контента и формирования поисковой выдачи. В работе [55] рассматриваются современные подходы краулинга и веб-скрапинга, а так же их применение. Отмечается их полезность в анализе социальных сетей, мониторинге цен и рыночного поведения, а также в агрегация новостей, данных, и создания датасетов.

Существует множество технологий, которые позволяют реализовать краулинг:

Scrapy — асинхронный фреймворк на Python для создания масштабируемых краулеров, активно используемый в промышленности и академических проектах [56].

BeautifulSoup — библиотека для парсинга HTML и XML, удобна при работе с небольшими статическими страницами [57].

Selenium — инструмент автоматизации браузеров, позволяющий имитировать действия пользователя на веб-страницах. Особенно полезен для работы с динамически генерируемым контентом [58].

Puppeteer — библиотека управления Chrome от Google, также подходит для высокоточного краулинга [59].

В рамках данной работы краулер будет реализован с использованием Selenium. Выбор этого инструмента обусловлен рядом факторов:

– Поддержка динамического контента: современные веб-страницы (в том числе социальные сети) часто загружают данные с помощью JavaScript. Selenium взаимодействует с реальным браузером, что позволяет получить доступ к этим данным,

– эмуляция пользовательских действий: Selenium позволяет программно кликать кнопки, заполнять поля, прокручивать ленту и т.д., что необходимо для обхода защитных механизмов и загрузки скрытого контента,

– работа с авторизацией: возможна реализация логина, управления сессиями и куки,

– гибкость и визуальный контроль: можно наблюдать за выполнением краулера в реальном времени, что облегчает отладку.

Таким образом, Selenium предоставляет возможности, необходимые для работы с защищёнными, динамическими и сложными по структуре сайтами, что делает его оптимальным выбором для целей данной работы.

2.4 Архитектура модели

Выбор архитектуры модели обусловлен необходимостью учёта контекстуальных связей в тексте, способности к обобщению на малых выборках и обеспечению интерпретируемости. Реализуемая архитектура основана на двунаправленной рекуррентной нейросети с памятью (BiLSTM) и включает в себя встроенный механизм внимания (attention), размещённый между выходом рекуррентного слоя и финальным классификатором. Такая комбинация обеспечивает как эффективное извлечение скрытых представлений из последовательности, так и их взвешенную агрегацию с акцентом на наиболее значимые фрагменты текста.

Основная архитектура двунаправленной LSTM была построена согласно работе в рамках которой исследовался BiLSTM-Attention механизм для эмоциональной классификации [60]. Модель строится следующим образом: На вход поступает последовательность токенов, полученная в результате предобработки текстов и трансформации их в векторное пространство. В качестве векторных представлений используются предварительно обученные эмбеддинги (например, FastText или Word2Vec, обученные на русскоязычных корпусах), представленные в виде матрицы. Эта матрица подаётся на слой BiLSTM, который обрабатывает текст в двух направлениях: слева направо и справа налево. В результате формируется тензор скрытых состояний размерности, содержащий контекстуальные признаки для каждого токена.

Следом за BiLSTM применяется механизм внимания, который рассчитывает вес значимости для каждого скрытого состояния. На основе этих весов формируется взвешенная сумма признаков, отражающая наиболее важные элементы текста.

Обработка метаданных должна проходить отдельно, так как способы работы с текстом и пользовательскими признаками отличаются. Метаданные пользователя будут подаваться в многослойный перцептрон для обработки признаков пользователя (пол, кол-во подписчиков, отношение к алкоголю и

др.). Многослойный перцептрон будет состоять из блоков, содержащих линейный слой (Linear), слой нормализации (BatchNorm1d), функцию активации (ReLU) и Dropout слой для регуляризации.

Полученные агрегированные векторы поступают в полно связный слой с активацией ReLU, который снижает размерность признаков и усиливает значимые аспекты. Завершает архитектуру выходной нейрон с функцией активации Sigmoid, обеспечивающий бинарную классификацию.

Концептуальная схема построенной модели представлена на рисунке 3. Она отражает полный путь обработки: от входных эмбеддингов слов (слева) и обработку метаданных справа, до классификационного выхода (внизу). Такая архитектура обеспечивает как устойчивость на малых выборках, так и интерпретируемость решений.

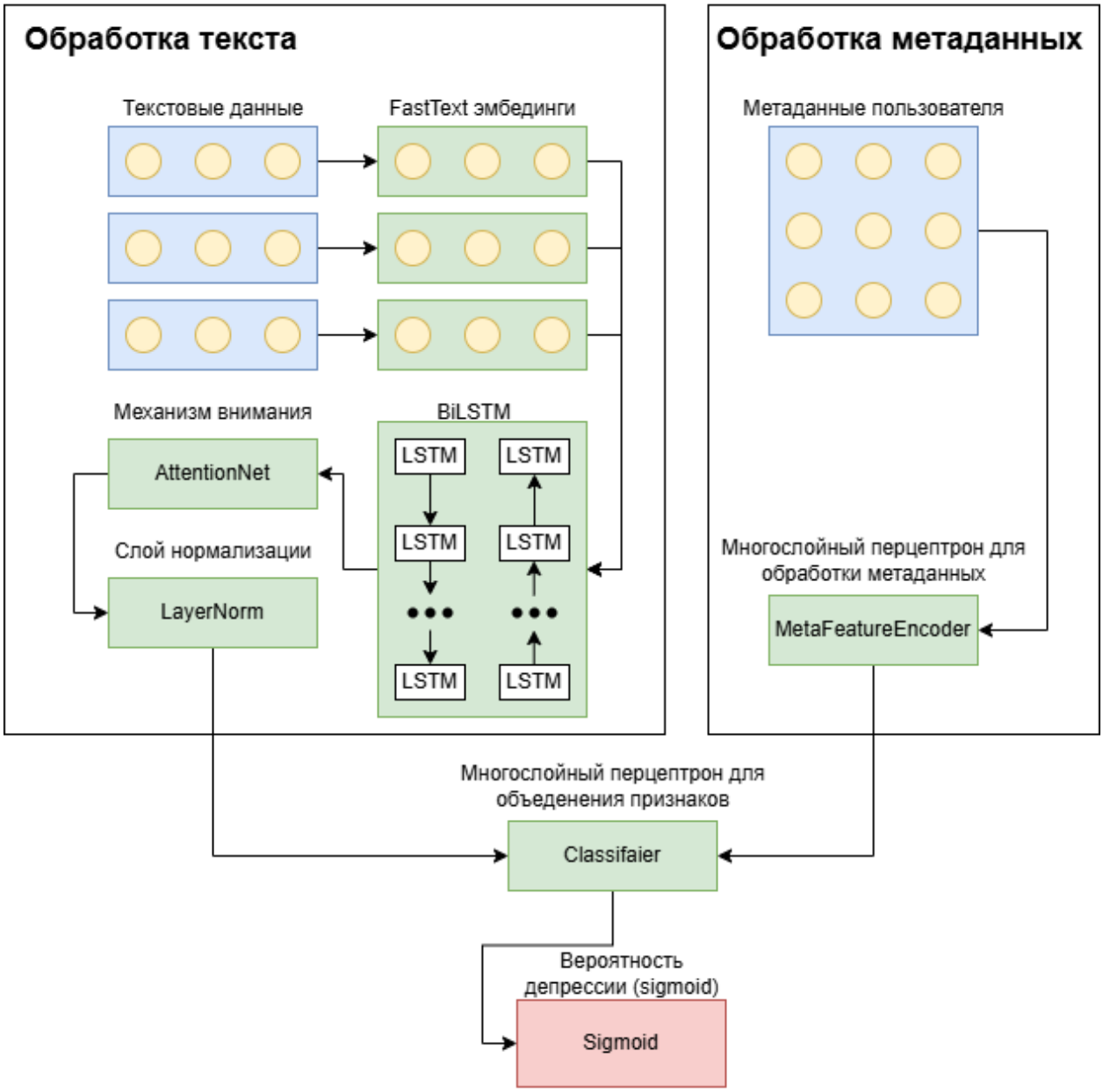

Рисунок 3 – Архитектура BiLSTM+Attention модели с использованием мета-признаков пользователей.

Таким образом, модель будет сочетать в себе преимущества рекуррентной обработки, внимания к ключевым словам и простоты классификации, оставаясь эффективной и интерпретируемой в условиях ограниченного объёма обучающих данных.

3. Практическая реализация проекта

3.1 Подготовка данных

В данном разделе решается третья задача исследования – сбор и подготовка данных, включая предобработку текстов, аугментацию и балансировку классов.

В рамках подготовки входных данных перед обучением модели нужно провести этап предобработки текстов. Для начала все посты пользователя объединяются в один, а затем процесс включает в себя несколько ключевых операций, направленных на устранение шума, нормализацию лексических форм и обеспечение совместимости с архитектурой нейросети. В частности, все тексты приводятся к нижнему регистру, удаляются гиперссылки, HTML-теги, специальные символы и знаки препинания, характерные социальным сетям упоминания и хештеги, не несущие семантической нагрузки. Удаление стоп-слов с использованием расширенного словаря (стандартные стоп-слова NLTK для русского языка [61] + специфические стоп-слова для социальных сетей), что помогло сильно снизить шум в данных.

Рисунок 4 - Облако слов до применения словаря стоп-слов.

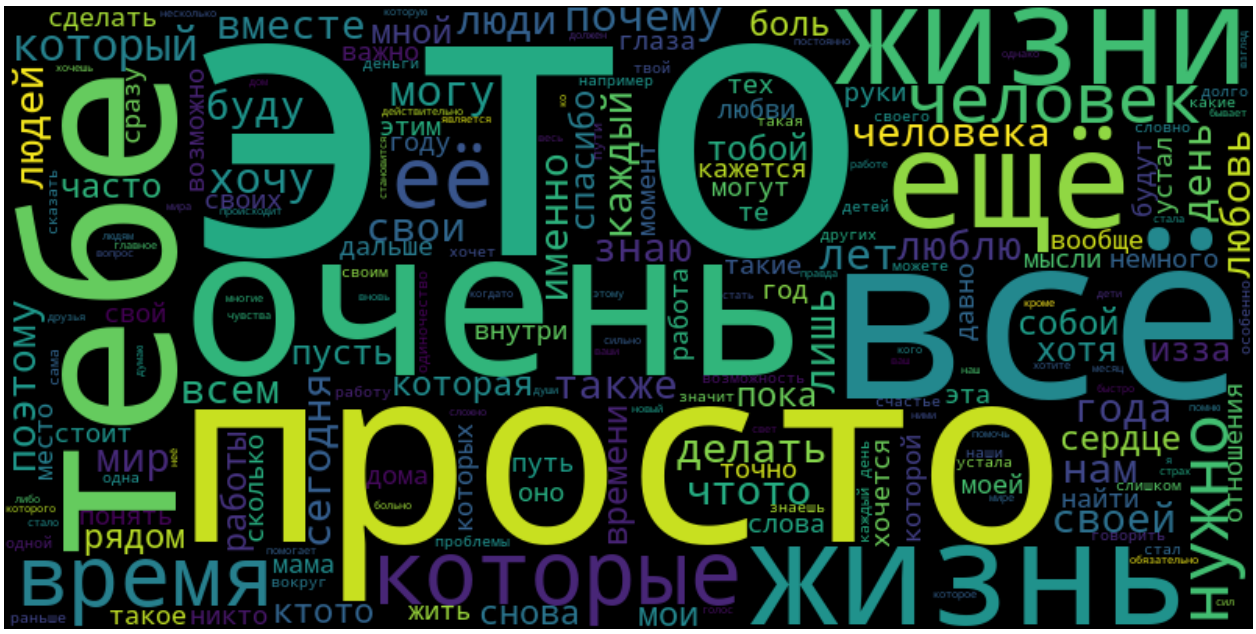

Рисунок 5 - Облако слов после применения словаря стоп-слов.

На представленных изображениях наглядно демонстрируется влияние этапа удаления стоп-слов на тексты данных депрессивного контекста. Первое облако слов (рисунок 4) содержит большое количество общих слов («это», «как», «не», «я»), которые не несут семантической ценности, создавая шум и усложняя анализ. После удаления стоп-слов (рисунок 5), включая как стандартные стоп-слова русского языка, так и специфические для социальных сетей, облако слов стало значительно информативнее. Чётко выделились ключевые слова, такие как «жизнь», «очень», «время», «боль», «устала», которые более точно отражают содержание текстов, характерных для депрессивного контекста.

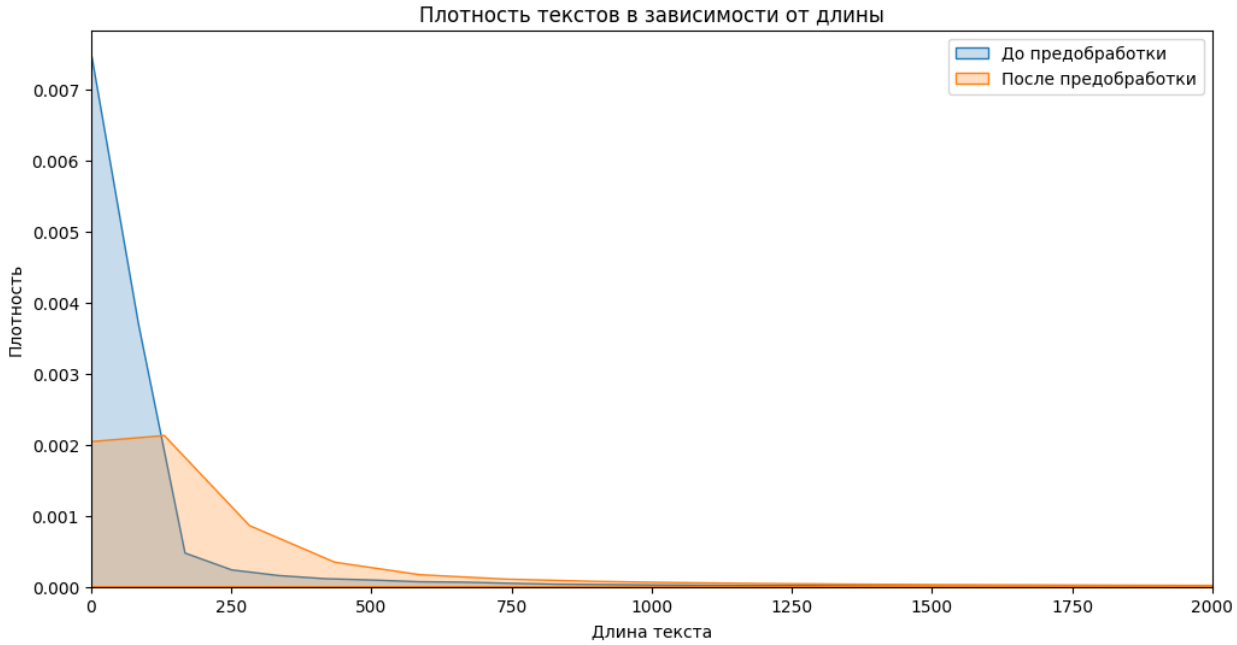

Рисунок 6 – Плотность текстов до и после предобработки.

График на рисунке 6 демонстрирует, что после предобработки текстов их средняя длина существенно снизилась, что является признаком удаления незначимой информации (стоп-слов, HTML-тегов, ссылок и других шумовых данных). Видно, что распределение текстов стало более узким, что указывает на приведение текстов к более компактному и информативному виду.

Далее осуществляется лемматизация текста с помощью библиотеки pymorphy3 [62], что особенно актуально при работе с русскоязычными текстами, обладающими высокой морфологической изменчивостью. Отдельно собираются мета признаки (пол, кол-во подписчиков, отношение к алкоголю и тп.), и подвергаются обработке выбросов позволяющей идентифицировать и соответствующим образом обрабатывать аномальные образцы, которые могут негативно влиять на обобщающую способность модели.

Для построения векторного представления текста используются предварительно обученные эмбеддинги FastText, так как они учитывают морфологию и устойчивы к орфографическим ошибкам и неформальной лексике, характерной для социальных сетей.

Особое внимание в процессе обучения модели уделяется проблеме несбалансированности классов. Анализ собранных данных показывает, что тексты с признаками депрессивного контекста составляют меньшинство по сравнению с нейтральными сообщениями, что потенциально может привести к смещению предсказаний модели в сторону доминирующего класса. Подобный дисбаланс отмечается и в ряде исследований, включая [7, 30, 37], и, как показывают авторы, требует специальных методов корректировки.

Первоначальное распределение включало всего 1327 положительных образцов (тексты с депрессивным контекстом) и 84587 отрицательных (тексты без признаков депрессии), что представляет соотношение примерно 1:64. Столь выраженный дисбаланс является критическим фактором, способным негативно повлиять на процесс обучения модели и привести к сильному смещению в сторону доминирующего класса.

Для уменьшения степени дисбаланса применяется аугментация положительных примеров, в результате которой добавлено 5218 синтетических образцов депрессивного контекста. Этот подход позволил увеличить размер миноритарного класса и получить более репрезентативную выборку для обучения. Аугментация текстовых данных выполняется с использованием двух методов - удаление случайных слов и перемешивание порядка некоторых слов. Удалялись и перемешивались около 10% от общего числа слов. Важно отметить, что текстовые примеры длиной менее 4 слов не подвергались аугментации, чтобы избежать искажения смысла. Для каждого положительного примера генерировалось до 4 дополнительных вариаций (2 с удалением слов и 2 с перемешиванием), что объясняет увеличение количества положительных примеров с 1327 до 6545.

Рисунок 7 – Количество депрессивных текстов до и после аугментации.

Для дальнейшей борьбы с проблемой несбалансированности классов был использован метод Partial SMOTE (Partial Synthetic Minority Over-sampling Technique). Данный подход позволяет частично выровнять распределение классов путем генерации синтетических примеров для миноритарного класса и одновременного сокращения числа образцов мажоритарного класса. В результате применения данной техники обучающая выборка была трансформирована и стала содержать 5236 положительных и 67669 отрицательных примеров, сохраняя соотношение классов примерно на уровне 1:13. Хотя дисбаланс все еще присутствует, его степень была скорректирована для создания более стабильных условий обучения модели.

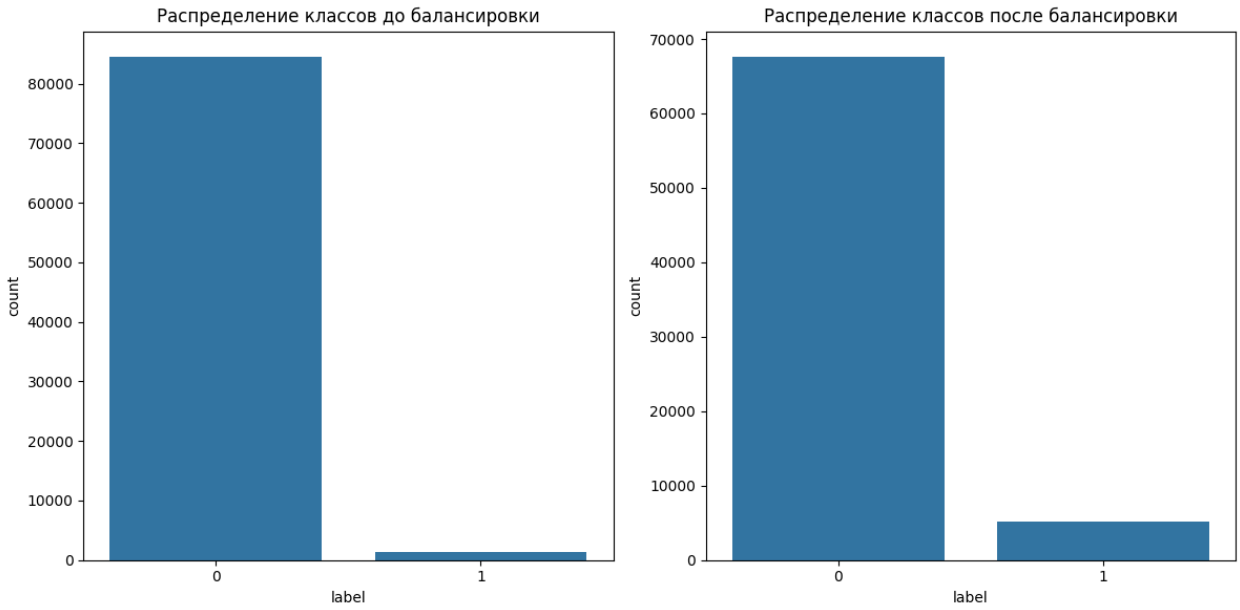

Рисунок 8 – Распределение классов до и после балансировки.

Для эффективного обучения нейронной сети текстовые данные были преобразованы в числовые представления с использованием предварительно обученных FastText эмбеддингов для русскоязычных текстов [63]. Данный подход позволяет эффективно преобразовывать слова в многомерные векторы, сохраняя их семантические свойства и взаимосвязи, что особенно важно при работе с лингвистическими нюансами, связанными с выражением депрессивных состояний.

Таким образом, на этапе подготовки данных был выполнен комплексный процесс очистки и нормализации текстов, включающий удаление стоп-слов, лемматизацию и обработку выбросов, что позволило существенно улучшить качество текстовых данных. Проблема несбалансированности классов решена с помощью методов аугментации и Partial SMOTE, что обеспечило представительность обучающей выборки. Данный этап создал основу для последующего эффективного обучения модели, позволяя учесть как текстовые, так и поведенческие характеристики пользователей.

3.2 BiLSTM+Attention

В рамках данного исследования была обучена нейросетевая модель двунаправленного LSTM с механизмом внимания и использованием метаданных. Процесс обучения модели осуществлялся со следующими гиперпараметрами:

- Размерность скрытого слоя: 128,
- количество слоев LSTM: 2,
- вероятность dropout: 0.5,
- скорость обучения: 0.001,
- количество эпох: 20,
- размер партии (batch size): 32.

Для улучшения сходимости и стабилизации процесса обучения применялись специальные методы инициализации весов, подобранные для конкретных типов слоев модели:

1) Xavier инициализация для линейных слоев — распределяет начальные значения весов равномерно в диапазоне, зависящем от размера входного и выходного слоя. Это позволяет поддерживать дисперсию сигналов при прямом и обратном проходе, что предотвращает затухание или взрыв градиентов в глубоких сетях. Смещения (bias) инициализировались нулями для уменьшения начального смещения модели,

2) Ортогональная инициализация для LSTM слоев — начальные матрицы весов формируются как ортогональные, что особенно важно для рекуррентных сетей. Ортогональные матрицы имеют собственные значения с модулем равным единице, что помогает предотвратить проблему исчезающих или взрывающихся градиентов при обработке длинных последовательностей. Применение этого метода значительно улучшает сходимость рекуррентных моделей,

3) Константная инициализация для слоев батч-нормализации — весовые коэффициенты инициализируются единицами, а смещения нулями, что обеспечивает «нейтральное» начальное поведение нормализованных слоев в начале обучения.

Такой подход к инициализации весов решает несколько ключевых проблем, возникающих при обучении глубоких нейронных сетей: ускоряет процесс сходимости, позволяя модели быстрее достичь оптимальных значений; значительно снижает вероятность застревания в плохих локальных минимумах функции потерь; предотвращает проблему исчезающих или взрывающихся градиентов, особенно актуальную для LSTM архитектур; обеспечивает более стабильную динамику обучения, особенно на ранних эпохах.

Учитывая существенный дисбаланс классов в данных (соотношение положительных к отрицательным примерам примерно 1:13 даже после аугментации), для эффективного обучения модели был применен механизм взвешивания классов. Этот подход реализован с помощью применения взвешенной функции потерь BCEWithLogitsLoss библиотеки PyTorch, с которой можно ознакомится в источнике [64]. Данная техника решает одну из ключевых проблем при работе с несбалансированными данными - предотвращает смещение модели в сторону преобладающего (отрицательного) класса, повышая чувствительность в его сторону. Комбинация взвешивания классов с другими техниками (аугментация, частичная балансировка методом Partial SMOTE) позволила эффективно справиться с проблемой дисбаланса классов и достичь высоких метрик на тестовой выборке даже для миноритарного класса.

В процессе обучения применялись следующие техники оптимизации:

1) Обрезка градиентов (градиентный клиппинг [65]) со значением 1.0, что способствовало стабилизации процесса обучения и предотвращало

проблему взрывающихся градиентов, характерную для рекуррентных нейронных сетей,

2) Планировщик скорости обучения (ReduceLROnPlateau [66]) для динамической корректировки параметра скорости обучения в зависимости от прогресса обучения, что позволяет более эффективно приближаться к оптимальным значениям весов модели,

3) Процедура ранней остановки (early stopping), которая прекращает обучение, если в течение трех последовательных эпох не наблюдается улучшение целевой метрики на валидационной выборке, что помогает предотвратить переобучение.

Более подробное описание компонентов архитектуры модели:

1) Текстовый поток обработки состоит из эмбеддингов FastText (300-мерные) для каждого слова в тексте, двунаправленной LSTM с 2 слоями (hidden_dim=128), механизма внимания для определения важности различных частей текста и нормализации выходного представления текста.

2) MetaFeaturesEncoder - многослойный перцептрон для обработки метаданных пользователя принимающий на вход метаданные: пол, количество подписчиков, отношение к алкоголю и курению, жизненные приоритеты. Состоит из 4 последовательных блока с уменьшающимися размерностями: 128 → 64 → 32 → 16. Каждый блок включает линейный слой, батч-нормализацию, ReLU и дропаут.

3) Классификатор - многослойный перцептрон для объединения признаков текстового представления и метаданных. Состоит из 4 последовательных блоков с уменьшающимися размерностями: 256 → 128 → 64 → 1. Каждый блок включает линейный слой, батч-нормализацию, ReLU и дропаут с уменьшающейся вероятностью.

Общее время обучения модели составило 13 часов 38 минут, что является приемлемым с учетом объема данных и сложности задачи. Анализ динамики обучения демонстрирует стабильное улучшение всех метрик на

протяжении 20 эпох тренировки. Потери на обучающей выборке уменьшились с 0.7882 на первой эпохе до 0.0735 на последней эпохе (рисунок 9). Точность (accuracy) увеличилась с 0.9495 до 0.9873. Метрика F1 выросла с 0.6937 до 0.9128. Значение AUC повысилось с 0.9599 до 0.9869.

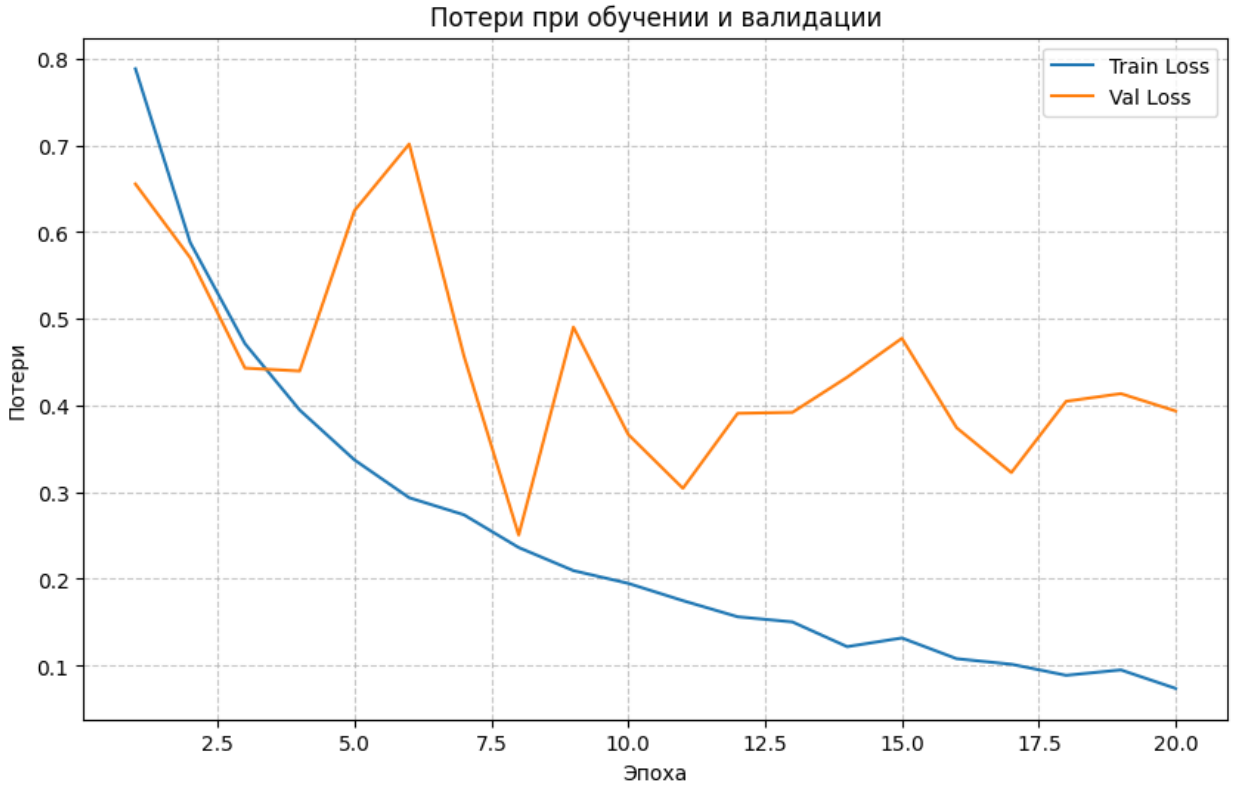

Рисунок 9 – Потери при обучении и валидации.

Такая динамика свидетельствует о высокой способности модели к обобщению и отсутствии явных признаков переобучения.

Финальная модель продемонстрировала высокие показатели эффективности на тестовой выборке:

Таблица 1. Результаты обучения модели.

Точность (accuracy)	Precision	Recall	F1	AUC	Loss
0.9873	0.8986	0.9274	0.9128	0.9869	0.3935

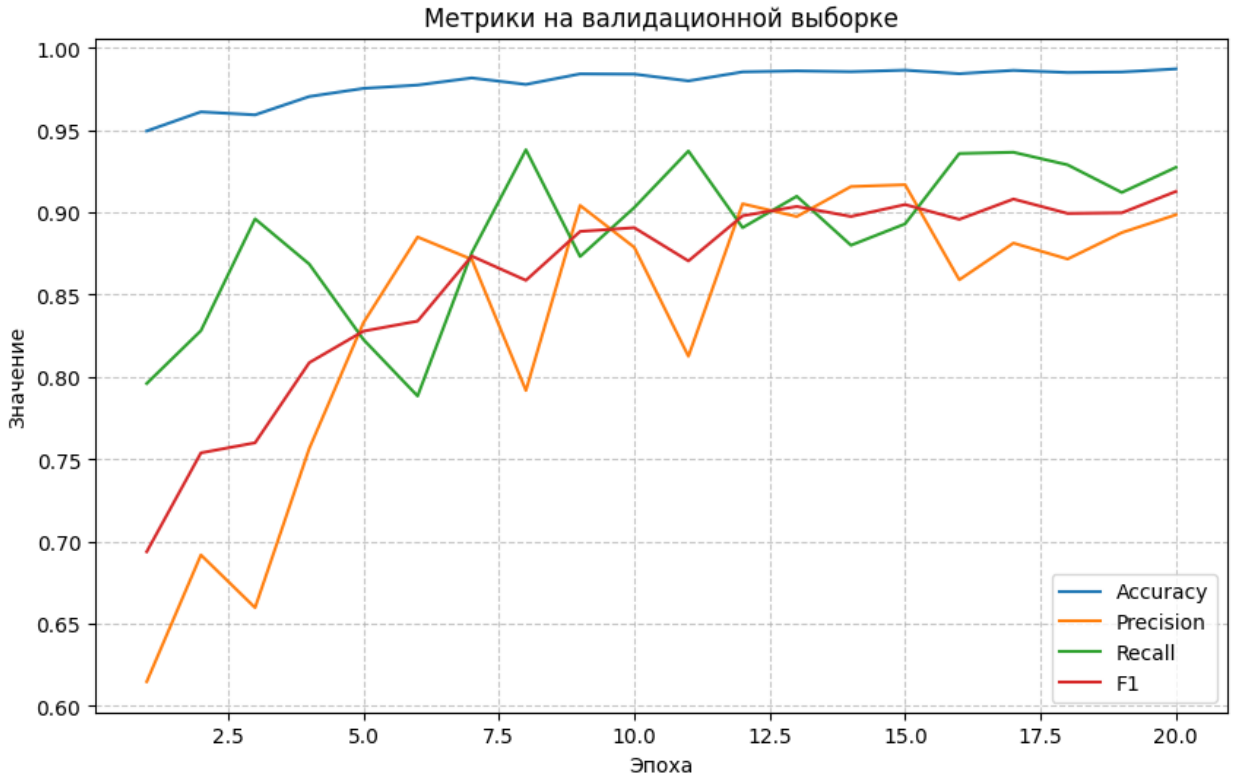

Рисунок 10 – Метрики на валидационной выборке.

Модель демонстрирует высокую эффективность при идентификации депрессивного контекста, о чем свидетельствуют высокие значения метрик precision и recall. Это особенно важно в контексте задачи выявления потенциально негативного контента, где ложноотрицательные результаты могут иметь серьезные последствия. Высокое значение AUC (0.9869) указывает на способность модели эффективно разделять классы, что является ключевым показателем в условиях несбалансированных данных.

Несмотря на высокие показатели эффективности, модель имеет ряд ограничений. Показатель precision (0.8986), хотя и высокий, указывает на наличие ложноположительных срабатываний, что может приводить к некорректной классификации нейтральных текстов как депрессивных. В контексте практического применения это может требовать дополнительной верификации результатов. Наблюдается некоторая нестабильность в динамике метрик в процессе обучения, что может указывать на чувствительность модели к особенностям отдельных партий данных и потенциальные проблемы с

локальными минимумами функции потерь. Архитектура LSTM, хотя и эффективна для обработки последовательных данных, имеет ограничения в способности улавливать долгосрочные зависимости в тексте, особенно в случаях с длинными последовательностями.

Разработанная модель представляет собой эффективный инструмент для анализа текстовых данных на предмет наличия депрессивного контекста. Достигнутые показатели эффективности свидетельствуют о высоком потенциале применения данной модели в практических задачах мониторинга контента и выявления потенциально проблемных текстов.

Дальнейшие исследования могут быть направлены на применение более сложных архитектур, таких как трансформеры, увеличение объема и разнообразия обучающих данных, а также на разработку более тонких методов регуляризации модели для повышения ее устойчивости и точности. Так же стоит отметить полезность незатронутого в этой работе контента социальных сетей – изображения, видео, аудио. Использование мультимодальных моделей позволит значительно увеличить качество учитывая особенности той же “ВКонтакте”, где большая часть контента представляется изображениями.

Более подробно ознакомится с подготовленной моделью, а также предобработкой и оптимизациями обучения можно в источнике [67].

3.3 Разработка краулера

Данный подраздел посвящен решению четвёртой задачи исследования – разработке краулера для автоматического сбора данных из социальной сети ВКонтакте.

Для сбора данных из социальной сети ВКонтакте был разработан собственный краулер, реализованный на языке Python с использованием библиотеки Selenium. Краулер позволяет автоматизировать процесс входа в аккаунт ВКонтакте, а также осуществлять сбор информации о профиле пользователя и его публикациях. Одной из его ключевых особенностей

является необходимость авторизации, обусловленная тем, что социальная сеть ВКонтакте ограничивает доступ к профилям пользователей без наличия учетной записи. В процессе авторизации краулер использует WebDriver Chrome через библиотеку Selenium, что позволяет имитировать действия пользователя, включая ввод логина и пароля. Для обеспечения безопасности учетной записи и в соответствии с настройками аккаунта, в котором активирована двухфакторная аутентификация (2FA), реализована автоматическая обработка этого процесса. В случае, если система запрашивает код подтверждения из SMS, краулер автоматически ожидает его ввода пользователем в консоль, что обеспечивает гибкость и возможность работы даже при наличии 2FA.

На рисунке 11 — схема работы краулера, отражающая этапы авторизации, сбора информации и обработки данных.

Рисунок 11 – Схема авторизации в краулере.

Краулер реализует обход защиты автоматизации, отключая атрибут «webdriver» в настройках браузера, что позволяет снизить вероятность блокировки. После успешной авторизации краулер переходит на страницу профиля пользователя и собирает данные, включая имя пользователя, идентификатор профиля, статус, количество подписчиков и другие характеристики. В зависимости от настроек конфиденциальности пользователя определяется публичный или закрытый статус профиля, и, в случае закрытого профиля, сбор данных ограничивается. Для сбора постов краулер автоматически прокручивает страницу, что позволяет загружать дополнительные публикации пользователя. Собранные данные включают текст постов, количество лайков, репостов, комментариев и просмотров. Эти данные сохраняются в формате JSON, что упрощает их дальнейшую обработку и анализ.

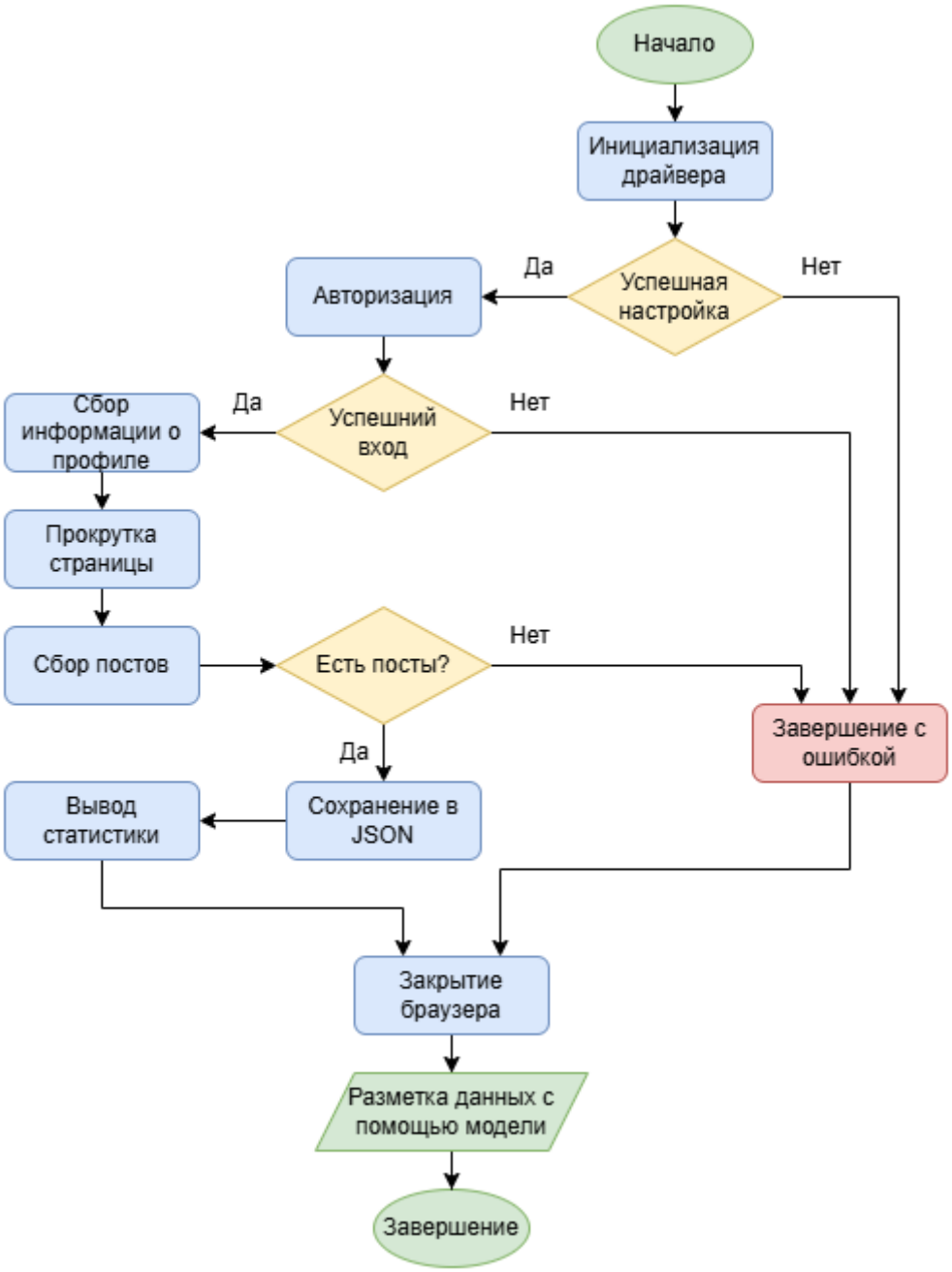

Рисунок 12 – Схема работы краулера.

После сбора данных краулер выполняет их разметку с помощью модели, полученной в разделе 3.2. К данным пользователей собранными краулером добавляются поля: label – в котором 1 значит, что пользователь помечен как депрессивный и 0 в противном случае; probability – вероятность депрессивности пользователя, выданная моделью.

Более подробно с работой краулера можно ознакомится в источнике [67].

ЗАКЛЮЧЕНИЕ

В ходе выполнения данной работы была решена комплексная научно-практическая задача по разработке ПО для автоматического выявления депрессивного контекста в текстах социальных сетей, что включало несколько ключевых этапов.

Во-первых, был проведен детальный анализ существующих исследований и методов выявления депрессивного контекста в текстах социальных сетей. Рассмотрены классические алгоритмы машинного обучения (логистическая регрессия, SVM, случайные леса), гибридные и мультимодальные подходы, а также современные нейросетевые архитектуры (RNN, LSTM, BiLSTM, трансформеры, включая BERT и SBERT). Это позволило выделить основные подходы и выбрать наилучший из них для решения поставленной задачи.

Во-вторых, была выбрана архитектура модели на основе двунаправленной рекуррентной нейронной сети (BiLSTM) с механизмом внимания (Attention). Этот выбор обоснован способностью модели учитывать долгосрочные и контекстуальные зависимости текста, что критически важно для анализа депрессивного контекста. Были определены и настроены гиперпараметры модели, обеспечивающие её устойчивость и высокую точность.

В-третьих, был создан и подготовлен уникальный датасет на основе текстов пользователей социальной сети ВКонтакте, включающий 85 914 пользователей и 206 805 постов. Разработана процедура эвристической разметки данных на основе лингвистических и психологических маркеров, что обеспечивает гибкость и адаптацию к русскоязычным текстам. Применена многоступенчатая система проверки (автоматическая фильтрация + ручная верификация), что повышает качество разметки. Данные прошли этапы предобработки, включая удаление шумовых данных, лемматизацию и балансировку классов. Для борьбы с дисбалансом данных применялись

методы аугментации и частичного SMOTE, что позволило значительно улучшить качество обучения модели.

В-четвёртых, была реализована и обучена модель BiLSTM+Attention, которая продемонстрировала высокие показатели эффективности на тестовой выборке: Accuracy — 0.9873, Precision — 0.8986, Recall — 0.9274, F1 — 0.9128, AUC — 0.9869. Эти результаты подтверждают гипотезу о том, что применение рекуррентных нейросетей (BiLSTM) с механизмом внимания позволяет с высокой точностью выявлять депрессивный контекст в текстах социальных сетей. Разработанная модель продемонстрировала высокую эффективность при выявлении депрессивного контекста в текстах социальной сети ВКонтакте, сохраняя при этом низкие требования к вычислительным ресурсам. В дополнение к текстам, модель учитывает мета-признаки пользователей (пол, количество подписчиков, отношение к алкоголю и т.д.). Благодаря данной архитектуре, модель способна точно классифицировать данные даже при ограниченном объеме обучающей выборки. В отличие от существующих решений, которые для достижения высоких показателей эффективности используют сложные трансформерные модели, требующих масштабных данных и высокопроизводительных вычислительных систем, предложенное решение обеспечивает оптимальный баланс между качеством классификации и доступностью для практического применения. Такой подход делает модель пригодной для использования в различных условиях, включая локальные системы с ограниченными ресурсами.

Практическая значимость данной работы заключается в создании программного решения, способного автоматически собирать данные из социальной сети ВКонтакте и определять депрессивный контекст текстов. Разработанная модель и краулер могут использоваться как в исследовательской деятельности, так и в рамках корпоративных систем мониторинга психоэмоционального состояния пользователей.

Дальнейшее развитие проекта может быть связано с интеграцией мультимодальных данных (изображения, видео, аудио), что позволит учитывать более широкий спектр информации и повысить точность модели. Такой подход позволит расширить применение системы на другие социальные платформы и улучшить её универсальность.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. МКБ 10 - Расстройства настроения [аффективные расстройства] (F30-F39). – URL: https://mkb-10.com/index.php?pid=4193(дата обращения: 07.04.2025). – Текст: электронный.

2. What Is Depression? – Depression (major depressive disorder) is a common and serious medical illness that negatively affects how you feel, the way you think and how you act. Fortunately, it is also treatable. – URL: https://www.psychiatry.org:443/patients-families/depression/what-is-depression(дата обращения: 23.03.2025). – Текст: электронный.

3. Кисельникова, Н. В. Выявление информативных параметров поведения пользователей социальной сети ВКонтакте как признаков депрессии / Н.В. Кисельникова, М.М. Данина, Е.В. Лаврова, Е.А. Куминская. – Текст : непосредственный. // Психология. Журнал Высшей школы экономики. 2020. Т. 17. № 1. – С. 73–88.

4. Солохов, Т. Д. Forecasting the depression with user data from Russian-language social network / Т.Д. Солохов, А.А. Кочкаров. – Текст : непосредственный. // МОДЕЛИРОВАНИЕ, ОПТИМИЗАЦИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ. 2024. Т. 12. № 2(45).

5. Pinto, S. J. Comprehensive review of depression detection techniques based on machine learning approach / S.J. Pinto, M. Parente. – Text : direct // Soft Computing. 2024. Vol. 28. № 17-18. – PP. 10701-10725.

6. Noh, S.-H. Analysis of Gradient Vanishing of RNNs and Performance Comparison / S.-H. Noh. – Text : direct // Information. 2021. Vol. 12. № 11. – PP. 442.

7. Sampath, K. Data Set Creation and Empirical Analysis for Detecting Signs of Depression from Social Media Postings / K. Sampath, T. Durairaj. – Text: direct // Computational Intelligence in Data Science / eds. L. Kalinathan [et al.]. – Cham: Springer International Publishing, 2022. – PP. 136-151.

8. Chen, G. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization / G. Chen, D. Ye, Z. Xing, J. Chen, E. Cambria. – Текст : непосредственный. // 2017 International Joint Conference on Neural Networks (IJCNN) 2017 International Joint Conference on Neural Networks (IJCNN). – Anchorage, AK, USA: IEEE, 2017. – С. 2377–2383.

9. Alghobiri, M. Guarding the Truth: Enhancing Fake Headline Detection using Transformer-Based Encoding and Deep Learning Methods / M. Alghobiri. – Текст : непосредственный. // International Journal of Open Information Technologies. 2024. Т. 12. Guarding the Truth. № 5. – С. 151–165.

10. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. – Текст : непосредственный. // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) NAACL-HLT 2019 / ред. J. Burstein, C. Doran, T. Solorio. – Minneapolis, Minnesota: Association for Computational Linguistics, 2019. BERT. – С. 4171–4186.

11. XGBoost Documentation — xgboost 3.0.0 documentation. – URL: https://xgboost.readthedocs.io/en/release_3.0.0/(дата обращения: 13.04.2025). – Текст: электронный.

12. Karaman, Y. A Comparative Analysis of SVM, LSTM and CNN-RNN Models for the BBC News Classification / Y. Karaman, F. Akdeniz, B.K. Savaş, Y. Becerikli. – Text : direct // Innovations in Smart Cities Applications Volume 6 : Lecture Notes in Networks and Systems / eds. M. Ben Ahmed [et al.]. – Cham: Springer International Publishing, 2023. Vol. 629. – PP. 473-483. – ISBN 978-3-031-26851-9.

13. AlSagri, H. S. Machine Learning-based Approach for Depression Detection in Twitter Using Content and Activity Features / H.S. AlSagri, M. Ykhlef. – Текст : непосредственный. // IEICE Transactions on Information and Systems. 2020. Т. E103.D. № 8. – С. 1825–1832.

14. Obagbuwa, I. C. Supervised machine learning models for depression sentiment analysis / I.C. Obagbuwa, S. Danster, O.C. Chibaya. – Text : direct // Frontiers in Artificial Intelligence. 2023. Vol. 6. – P. 1230649.

15. Liu, D. Detecting and Measuring Depression on Social Media Using a Machine Learning Approach: Systematic Review / D. Liu, X.L. Feng, F. Ahmed, M. Shahid, J. Guo. – Текст : непосредственный. // JMIR Mental Health. 2022. Т. 9. Detecting and Measuring Depression on Social Media Using a Machine Learning Approach. № 3. – С. e27244.

16. Salas-Zárate, R. Detecting Depression Signs on Social Media: A Systematic Literature Review / R. Salas-Zárate, G. Alor-Hernández, M. del P. Salas-Zárate, M.A. Paredes-Valverde, M. Bustos-López, J.L. Sánchez-Cervantes. – Text : direct // Healthcare. 2022. Vol. 10. Detecting Depression Signs on Social Media. № 2. – PP. 291.

17. Ding, Y. A Depression Recognition Method for College Students Using Deep Integrated Support Vector Algorithm / Y. Ding, X. Chen, Q. Fu, S. Zhong. – Text : direct // IEEE Access. 2020. Vol. 8. – PP. 75616-75629.

18. Александровна, З. А. Обнаружение депрессии среди пользователей социальной сети с использованием методов машинного обучения / З.А. Александровна, М.А. Иванович. – Текст : непосредственный. // Computational nanotechnology. 2023. Т. 10. № 4. – С. 16–22.

19. Smetanin, S. The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives / S. Smetanin. – Text : direct // IEEE Access. 2020. Vol. 8. The Applications of Sentiment Analysis for Russian Language Texts. – PP. 110693-110719.

20. Bilal, S. Суммаризация текста на арабском языке с использованием трехуровневой архитектуры двунаправленной долговременной краткосрочной памяти (BiLSTM) / S. Bilal. – Text : direct // Vestnik of Russian New University. Series “Complex systems: models, analysis, management”. 2024. № 1. – P. 75-85.

21. Bao, E. Explainable depression symptom detection in social media / E. Bao, A. Pérez, J. Parapar. – Text : direct // Health Information Science and Systems. 2024. Vol. 12. № 1. – PP. 47.

22. Tahir, W. B. Depression Detection in Social Media: A Comprehensive Review of Machine Learning and Deep Learning Techniques / W.B. Tahir, S. Khalid, S. Almutairi, M. Abohashrh, S.A. Memon, J. Khan. – Текст : непосредственный. // IEEE Access. 2025. Т. 13. Depression Detection in Social Media. – С. 12789–12818.

23. Babu, N. V. Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review / N.V. Babu, E.G.M. Kanaga. – Text : direct // SN Computer Science. 2022. Vol. 3. Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence. № 1. – PP. 74.

24. Zogan, H. DepressionNet: Learning Multi-modalities with User Post Summarization for Depression Detection on Social Media / H. Zogan, I. Razzak, S. Jameel, G. Xu. – Text : direct // Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. – Virtual Event Canada: ACM, 2021. DepressionNet. – PP. 133-142.

25. Александрович, Б. А. ОПРЕДЕЛЕНИЕ ПСИХИЧЕСКОГО СОСТОЯНИЯ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНОЙ СЕТИ REDDIT НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ / Б.А. Александрович, Ш.Я. Джитендер, К.И. Витальевич, Ф.Е. Владимировна, К.А. Владимирович, У.И. Александрович. – Текст : непосредственный. // Информационно-управляющие системы. 2022. № 1 (116). – С. 8–18.

26. Arif, M. Mental Illness Classification on Social Media Texts Using Deep Learning and Transfer Learning / M. Arif, I. Ameer, N. Bölücü, G. Sidorov, A. Gelbukh, V. Elangovan, M. Arif, I. Ameer, N. Bölücü, G. Sidorov, A. Gelbukh, V.

Elangovan. – Text : direct // Computación y Sistemas. 2024. Vol. 28. № 2. – P. 451-464.

27. Cao, Y. Machine Learning Approaches for Depression Detection on Social Media: A Systematic Review of Biases and Methodological Challenges. / Y. Cao, J. Dai, Z. Wang, Y. Zhang, X. Shen, Y. Liu, Y. Tian. – Text : direct // Journal of Behavioral Data Science. 2025. Vol. 5. Machine Learning Approaches for Depression Detection on Social Media. № 1. – P. 1-36.

28. Liu, Y. Learning Natural Language Inference using Bidirectional LSTM model and Inner-Attention / Y. Liu, C. Sun, L. Lin, X. Wang. – Текст : непосредственный. 2016.

29. Кузнецов, Р. С. ПРОГНОЗИРОВАНИЕ БИРЖЕВЫХ КОТИРОВОК AMAZON INC. С ИСПОЛЬЗОВАНИЕМ BILSTM-ATTENTION НЕЙРОННОЙ СЕТИ / Р.С. Кузнецов. – Текст : непосредственный. // Экономика и бизнес: теория и практика. 2023. № 10-2 (104). – С. 19–23.

30. Chen, Z. Detecting Reddit Users with Depression Using a Hybrid Neural Network SBERT-CNN / Z. Chen, R. Yang, S. Fu, N. Zong, H. Liu, M. Huang. – Текст : непосредственный. // 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI) / arXiv:2302.02759 [cs]. 2023. – С. 193–199.

31. Kour, H. An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM / H. Kour, M.K. Gupta. – Текст : непосредственный. // Multimedia Tools and Applications. 2022. Т. 81. № 17. – С. 23649–23685.

32. Ezen-Can, A. A Comparison of LSTM and BERT for Small Corpus / A. Ezen-Can. // *arXiv preprint arXiv:2009.05451*. 2020. – URL: http://arxiv.org/abs/2009.05451(дата обращения: 02.03.2025) – Текст: электронный.

33. Гольдберг Й. Нейросетевые методы в обработке естественного языка / пер. с анг. А. А. Слинкина. – Москва: ДМК Пресс, 2019. 282с. – ISBN 978-5-9706-0754-1. – Текст : непосредственный.

34. Vajrobol, V. Depression detection in Thai language posts based on attentive network models / V. Vajrobol, U. Shukla, A. Pundir, S. Singh, G. Saxena. – Text : direct. // WNLPe-Health@ICON. 2022.

35. Triantafyllopoulos, I. Depression detection in social media posts using affective and social norm features / I. Triantafyllopoulos, G. Paraskevopoulos, A. Potamianos. // *arXiv preprint arXiv:2303.14279*. 2023. – URL: https://arxiv.org/abs/2303.14279(дата обращения: 02.03.2025) – Текст: электронный.

36. Выявление информативных параметров поведения пользователей ВКонтакте как признаков депрессии //Психологическая газета. – Характер использования социальных сетей и определенные онлайн-действия (такие как частота обновлений, постов, добавление в друзья бывших партнеров или подписки на незнакомых людей) могут быть важными маркерами симптомов депрессии. – URL: https://psy.su/feed/9376/(дата обращения: 02.03.2025). – Текст: электронный.

37. Солохов, Т. Д. Выявление признаков депрессии на основе пользовательских данных из социальных сетей с помощью нейронных сетей: Detection of depression features with user data from social network using neural network / Т.Д. Солохов, А.А. Кочкаров. – Текст : непосредственный. // МОДЕЛИРОВАНИЕ, ОПТИМИЗАЦИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ. 2025. Т. 13. Выявление признаков депрессии на основе пользовательских данных из социальных сетей с помощью нейронных сетей. № 1(48).

38. Smetanin, S. Deep transfer learning baselines for sentiment analysis in Russian / S. Smetanin, M. Komarov. – Text : direct // Information Processing & Management. 2021. Vol. 58. № 3. – PP. 102484.

39. National research nuclear University MEPhI, Moscow, Russia. BiLSTM-based Approach to the Natural Language Text Dependencies Analysis / National

research nuclear University MEPhI, Moscow, Russia, A. Chernyshov. – Текст : непосредственный. // Information and Innovations. 2019. Т. 14. № 1. – С. 44–47.

40. Richter, T. Machine Learning-Based Behavioral Diagnostic Tools for Depression: Advances, Challenges, and Future Directions / T. Richter, B. Fishbain, G. Richter-Levin, H. Okon-Singer. – Text : direct // Journal of Personalized Medicine. 2021. Vol. 11. Machine Learning-Based Behavioral Diagnostic Tools for Depression. № 10. – PP. 957.

41. PhenomNet: Bridging Phenotype-Genotype Gap: A CNN-LSTM Based Automatic Plant Root Anatomization System. – PDF | This research will explore the phenotype-genotype gap by bringing two very diverse technologies together to predict plant characteristics.... | Find, read and cite all the research you need on ResearchGate. – URL: https://www.researchgate.net/publication/341131167_PhenomNet_Bridging_Phenotype-Genotype_Gap_A_CNN-LSTM_Based_Automatic_Plant_Root_Anatomization_System(дата обращения: 20.03.2025). – Текст: электронный.

42. Figure 2: RNN v/s LSTM. a: RNNs use their internal state (memory) to... – Download scientific diagram | RNN v/s LSTM. a: RNNs use their internal state (memory) to process sequences of inputs, b: Long Short-Term Memory (LSTM) network is a varient of RNN, with addtional long term memory to remember past data. from publication: PhenomNet: Bridging Phenotype-Genotype Gap: A CNN-LSTM Based Automatic Plant Root Anatomization System | This research will explore the phenotype-genotype gap by bringing two very diverse technologies together to predict plant characteristics. Currently, there are several studies and tools available for plant phenotype and genotype analysis. However, there is no existing single... | plant roots, Bridges and Systems | ResearchGate, the professional network for scientists. – URL: https://www.researchgate.net/figure/RNN-v-s-LSTM-a-RNNs-use-their-internal-state-memory-to-process-sequences-of-inputs_fig1_341131167(дата обращения: 27.05.2025). – Текст: электронный.

43. Farruque, N. Depression symptoms modelling from social media text: an LLM driven semi-supervised learning approach / N. Farruque, R. Goebel, S. Sivapalan, O.R. Zaïane. – Text : direct // Language Resources and Evaluation. 2024. Vol. 58. Depression symptoms modelling from social media text. № 3. – PP. 1013-1041.

44. VK для разработчиков. – Мощная платформа для ваших проектов. Разрабатывайте приложения и используйте все возможности ВКонтакте в вашем бизнесе. – URL: https://dev.vk.com/ru(дата обращения: 17.03.2025). – Текст: электронный.

45. PyPI · The Python Package Index. – The Python Package Index (PyPI) is a repository of software for the Python programming language. – URL: https://pypi.org/(дата обращения: 30.04.2025). – Текст: электронный.

46. vk-parser: Модульный парсер данных пользователей из социальной сети ВКонтакте.vk-parser. – URL: https://pypi.org/project/vk-parser/(дата обращения: 30.04.2025). – Текст: электронный.

47. Правила | Оферта на оказание услуг по привлечению пользователей к Сайту | VK для разработчиков. – apps-offer. – URL: https://dev.vk.com/ru/apps-offer(дата обращения: 30.03.2025). – Текст: электронный.

48. Higher School of Economics. Ethical and legal aspects of social media data usage / Higher School of Economics, I.A. Shcheglova. – Текст : непосредственный. // Vestnik Tomskogo gosudarstvennogo universiteta. 2018. № 431. – С. 81–87.

49. Кодекс этики использования данных. – URL: https://rubda.ru/wp-content/uploads/2021/05/kodeks-etiki.pdf(дата обращения: 30.03.2025). – Текст: электронный.

50. Benton, A. Ethical Research Protocols for Social Media Health Research / A. Benton, G. Coppersmith, M. Dredze. – Текст : непосредственный. // Proceedings of the First ACL Workshop on Ethics in Natural Language Processing

EthNLP 2017 / ред. D. Hovy [и др.]. – Valencia, Spain: Association for Computational Linguistics, 2017. – PP. 94–102.

51. Owen, D. AI for Analyzing Mental Health Disorders Among Social Media Users: Quarter-Century Narrative Review of Progress and Challenges / D. Owen, A.J. Lynham, S.E. Smart, A.F. Pardiñas, J. Camacho Collados. – Text : direct // Journal of Medical Internet Research. 2024. Vol. 26. AI for Analyzing Mental Health Disorders Among Social Media Users. – PP. e59225.

52. Vayena, E. Health Research with Big Data: Time for Systemic Oversight / E. Vayena, A. Blasimme. – Text : direct // Journal of Law, Medicine & Ethics. 2018. Vol. 46. Health Research with Big Data. № 1. – PP. 119-129.

53. crawler_of_depressive_context/cleaner at main · johnneon/crawler_of_depressive_context. – URL: https://github.com/johnneon/crawler_of_depressive_context/tree/main/cleaner#%D0%BA%D0%B0%D1%82%D0%B5%D0%B3%D0%BE%D1%80%D0%B8%D0%B8-%D0%BA%D0%BB%D1%8E%D1%87%D0%B5%D0%B2%D1%8B%D1%85-%D1%81%D0%BB%D0%BE%D0%B2(дата обращения: 10.05.2025). – Текст: электронный.

54. Olston, C. Web Crawling / C. Olston, M. Najork. – Text : direct // Foundations and Trends® in Information Retrieval. 2010. Vol. 4. № 3. – P. 175-246.

55. Khder, M. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application / M. Khder. – Text : direct // International Journal of Advances in Soft Computing and its Applications. 2021. Vol. 13. Web Scraping or Web Crawling. № 3. – P. 145-168.

56. Scrapy | A Fast and Powerful Scraping and Web Crawling Framework. – URL: https://scrapy.org/(дата обращения: 10.04.2025). – Текст: электронный.

57. Beautiful Soup Documentation — Beautiful Soup 4.13.0 documentation. – URL: https://www.crummy.com/software/BeautifulSoup/bs4/doc/(дата обращения: 10.04.2025). – Текст: электронный.

58. Selenium. – Selenium automates browsers. That’s it! What you do with that power is entirely up to you.Primarily it is for automating web applications for testing purposes, but is certainly not limited to just that. Boring web-based administration tasks can (and should) also be automated as well. Getting Started Selenium WebDriver Selenium WebDriver If you want to create robust, browser-based regression automation suites and tests, scale and distribute scripts across many environments, then you want to use Selenium WebDriver, a collection of language specific bindings to drive a browser - the way it is meant to be driven. – URL: https://www.selenium.dev/(дата обращения: 10.04.2025). – Текст: электронный.

59. Puppeteer | Puppeteer. – build. – URL: https://pptr.dev/(дата обращения: 10.04.2025). – Текст: электронный.

60. Zhou, Q. NLP at IEST 2018: BiLSTM-Attention and LSTM-Attention via Soft Voting in Emotion Classification / Q. Zhou, H. Wu. – Text : direct // Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. – Brussels, Belgium: Association for Computational Linguistics, 2018. NLP at IEST 2018. – P. 189-194.

61. node-nltk-stopwords/data/stopwords/russian at master · xiamx/node-nltk-stopwords. – A node module exposing nltk stopwords corpora and provide utility functions for removing stopwords - xiamx/node-nltk-stopwords. – URL: https://github.com/xiamx/node-nltk-stopwords/blob/master/data/stopwords/russian(дата обращения: 09.05.2025). – Текст: электронный.

62. pymorphy3: Morphological analyzer (POS tagger + inflection engine) for Russian language.pymorphy3. – URL: https://pypi.org/project/pymorphy3/(дата обращения: 10.05.2025). – Текст: электронный.

63. Word vectors for 157 languages · fastText. – We distribute pre-trained word vectors for 157 languages, trained on [*Common

Crawl*](http://commoncrawl.org/) and [*Wikipedia*](https://www.wikipedia.org) using fastText. – URL: https://fasttext.cc/index.html(дата обращения: 09.05.2025). – Текст: электронный.

64. BCEWithLogitsLoss — PyTorch 2.7 documentation. – URL: https://docs.pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html(дата обращения: 10.05.2025). – Текст: электронный.

65. torch.nn.utils.clip_grad_norm_ — PyTorch 2.7 documentation. – URL: https://docs.pytorch.org/docs/stable/generated/torch.nn.utils.clip_grad_norm_.html #torch-nn-utils-clip-grad-norm(дата обращения: 10.05.2025). – Текст: электронный.

66. ReduceLROnPlateau — PyTorch 2.7 documentation. – URL: https://docs.pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLR OnPlateau.html#reducelronplateau(дата обращения: 10.05.2025). – Текст: электронный.

67. johnneon/crawler_of_depressive_context. – Contribute to johnneon/crawler_of_depressive_context development by creating an account on GitHub. – URL: https://github.com/johnneon/crawler_of_depressive_context(дата обращения: 10.05.2025). – Текст: электронный.

ПРИЛОЖЕНИЕ А

Сравнительная таблица моделей.

Исследование	Модель	F1 (best)	Accuracy	Датасет
Arif и др. (2024)	BiLSTM, CNN, BERT, RoBERTa, AlBERT	до 0.89 (RoBERTa)	91%	Reddit posts (16,703 записей)
Bao et al. (2024)	WT5, WBART, GPT-3.5, Vicuna, MentalLLaMA	0.96	—	Reddit (PsySym, BDI-Sen, DepreSym, >1,700 симптомных постов + контрольные)
Kour & Gupta (2022)	CNN-biLSTM (гибрид)	0. 94	94.28%	Twitter, точный объём не указан
Ding et al. (2020)	DISVM (AdaBoost + SVM)	—	~86%	Sina Weibo (1000 студентов, данные за 2015–2017 гг.)
Chen et al. (2023)	SBERT + CNN (гибрид)	0.86	86%	Reddit (SMHD: 209,188 постов от 1,316 пользователей с депрессией + контрольная группа)
Vajrobol et al. (2024)	XLM-RoBERTa, M-BERT, Bi-GRU, CNN, LSTM, SVM и др.	0.8016 (XLM-R)	79.12% (XLM-R)	Thai Depression Dataset: 17,116 + 16,320 постов
Zogan et al., (2021)	DepressionNet (CNN + BiGRU + attention + summarization)	0.912	90%	Twitter (Depressed — 2159 пользователей, 447,856 твитов; Non-depressed — 2049 пользователей, 1,349,447 твитов)
Александрович и др., 2022	SVM (One-vs-Rest), fastText, CNN	0.796 (OvR)	80%(OvR)	Reddit (собран через Pushshift API, ~120 тыс. постов)

Продолжение таблицы 1.

Александровна и Иванович, 2023	XGBoost, SVM, Random Forest, Logistic Regression	—	77% (XGBoost)	ВКонтакте: 270 депрессивных, 231 контрольный, 5962 поста
Triantafyllopoulos et al., 2023	BERT + BiGRU + Attention + Emotion detector (B+E+P+M)	0.95 (Pirina), 0.93 (RSDD)	93.87% (Pirina)	Reddit (Pirina — 1,841 пост, RSDD — 28 млн постов, ~6,000 пользователей)
Чернышов А., 2019	BiLSTM	~0.74	~75%	Русскоязычные тексты, объём не указан
Солохов и Кочкаров, 2024	Логистическая регрессия, Random Forest, Gradient Boosting	0.92 (лог. регрессия), 0.98 (лес, бустинг)	92% (лог. регрессия), 98% (лес, бустинг)	ВКонтакте (584,000 постов от 49,600 пользователей, 550 с признаками депрессии)
Liu et al., 2016	BiLSTM + Inner Attention	—	85.0%	SNLI, 570 тыс. пар предложений
AlSagri и Ykhlef (2020)	SVM (linear, radial), Naive Bayes, Decision Tree	0.79	82.5%	Twitter (~300,000 твитов, 111 пользователей)
Obagbuwa и др. (2023)	Logistic Regression, SVM, XGB, Random Forest	—	96.3% (Logistic Regression), 96.2% (SVM)	Twitter (объединённый набор из 5 Kaggle-датасетов, включая 1.6 млн твитов)
Зоткина и Мартышкин (2023)	Logistic Regression, Random Forest, SVM, XGBoost	0.77 (XGBoost)	77%(XGBoost)	ВКонтакте (270 пользователей с депрессией, 2086 постов; 231 контрольный, 3876 постов)

ПРИЛОЖЕНИЕ Б

Пример готовых полученных данных.


```
{
  "user_id": 0,
  "sex": 2,
  "city": "Saint Petersburg",
  "followers_count": 11112,
  "alcohol": 0,
  "smoking": 0,
  "life_main": 0,
  "people_main": 0,
  "status": "создаю будущее",
  "posts": [
    {
      "text": "на videotech23 представил эксклюзивный keynote последние кадры реальности или...",
      "date": 1700817537,
      "likes": 186,
      "comments": 0,
      "reposts": 10,
      "views": 9359
    },
    {
      "text": "привет друзья мне выпала честь стать номинантом премии highload 2023 за первое...",
      "date": 1697200110,
      "likes": 235,
      "comments": 0,
      "reposts": 14,
      "views": 9000
    }
  ],
  "label": 0
},
{
  "user_id": 1,
  "sex": 2,
  "city": "Saint Petersburg",
  "followers_count": 3071,
  "alcohol": 2,
  "smoking": 1,
  "life_main": 0,
  "people_main": 0,
  "status": "6",
  "posts": [
    {
      "text": "уже второй месяц я страдаю, испытываю одиночество и у меня депрессия...",
      "date": 1693482203,
      "likes": 51,
      "comments": 1,
      "reposts": 2,
      "views": 1956
    }
  ],
  "label": 1
}
```