«Проект NLP по Ментальным Проблемам людей»

МФТИ

Муниров Султан

Весна 2025

Содержание

Diverse Perspectives, Divergent Models: Cross-Cultural Evaluation of Depression Detection on Twitter

Авторы: Nuredin Ali, Charles Chuankai Zhang , Ned Mayo , Stevie Chancellor

Аннотация: Данные социальных сетей используются для выявления пользователей с психическими расстройствами, такими как депрессия. В данной работе оценивается обобщаемость эталонных наборов данных для построения ИИ-моделей на кросс-культурных данных из Twitter. Результаты показывают, что модели обнаружения депрессии не обладают глобальной обобщаемостью, демонстрируя значительный разрыв в производительности между пользователями из стран Глобального Севера и Юга. Предобученные языковые модели показывают лучшие результаты, но сохраняют культурные смещения. Предложены рекомендации для улучшения репрезентативности данных.

Наборы данных

- **CLPsych**: 327 пользователей с депрессией, 570 контрольных (Coppersmith et al., 2015).
- **Multi-Task Learning (MTL)**: 1520 пользователей с депрессией, 1520 контрольных (Shen et al., 2017).
- **Пользовательский набор**: 267 пользователей с депрессией из 7 стран, проверенных вручную.

Модели

- **Logistic Regression**: Признаки TF-IDF, L2-регуляризация.
- **MentalLongformer**: Предобученная трансформерная модель, дообученная для обнаружения депрессии.

Метрики оценки

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

Производительность по странам

Training Data	Australia		Nigeria		South Africa		Philippines		India		UK		US	
	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1
CLPsych	0.61	0.63	0.13	0.23	0.45	0.53	0.39	0.46	0.10	0.19	0.53	0.61	0.53	**0.66**
Multi Task Learning	0.53	0.60	0.13	0.23	0.28	0.36	0.08	0.15	0.26	0.35	0.84	**0.69**	0.75	0.61

Table 3: F1 scores of *Logistic Regression* trained on CLPsych and Multi-Task Learning datasets.

Training Data	Australia		Nigeria		South Africa		Philippines		India		UK		US	
	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1
CLPsych	0.64	**0.72**	0.06	0.12	0.2	0.32	0.18	0.3	0.15	0.27	0.37	0.5	0.42	0.56
Multi Task Learning	0.93	0.69	0.33	0.45	0.71	0.67	0.31	0.43	0.68	0.7	0.95	**0.72**	0.84	0.64

Table 4: F1 scores of *MentalLongformer* trained on CLPsych and Multi-Task Learning datasets.

Рис. 1: Оценки F1 для разных стран

Global North vs. Global South

Таблица 1: Оценки F1 по регионам

Model	Global North	Global South
Logistic Regression	0.63	0.26
MentalLongformer	0.68	0.56

Суммаризация

- Низкая производительность на пользователях из Глобального Юга связана с лингвистическими и культурными различиями.
- Смешение языков (например, английский + тагальский) усложняет обнаружение депрессии.
- Рекомендации: увеличение объёма данных, улучшение обработки смешения языков, валидация на других платформах.

Towards Interpretable Mental Health Analysis with Large Language Models

Основные положения

Статья посвящена исследованию возможностей больших языковых моделей (LLMs) в анализе психического здоровья с акцентом на интерпретируемость. Авторы фокусируются на трёх ключевых аспектах:

- Оценка производительности LLMs (ChatGPT, LLaMA, InstructGPT-3) на 11 датасетах для 5 задач: бинарная/многоклассовая классификация психических состояний, определение причин расстройств, распознавание эмоций в диалогах.
- Исследование влияния различных стратегий подсказок (prompt engineering), включая эмоционально усиленные подсказки и few-shot обучение.
- Генерация и оценка объяснений моделей для повышения прозрачности решений.

Методология

- **Модели**: Сравнение ChatGPT, LLaMA-7B/13B, InstructGPT-3 с традиционными нейросетевыми архитектурами (CNN, GRU) и специализированными моделями (MentalBERT, MentalRoBERTa).
- **Стратегии подсказок**:
  - Zero-shot и Chain-of-Thought (CoT) prompting.
  - Эмоционально усиленные подсказки с использованием лексиконов VADER и NRC EmoLex.
  - Few-shot обучение с примерами, написанными экспертами.
- **Оценка объяснений**: Человеческая оценка по критериям беглости, надёжности, полноты (Fleiss' Kappa $> 0.21 > 0.21$) и автоматическая оценка с использованием BLEU, ROUGE, BARTScore.

Model	DR		CLPsych15		Dreaddit		T-SID		SAD		CAMS	
	Rec.	F1	Rec.	F1	Rec.	F1	Rec.	F1	Rec.	F1	Rec.	F1
Supervised Methods												
CNN	80.54	79.78	51.67	40.28	65.31	64.99	71.88	71.77	39.71	38.45	36.26	34.63
GRU	61.72	62.13	50.00	46.76	55.52	54.92	67.50	67.35	35.91	34.79	34.19	29.33
BiLSTM_Att	79.56	79.41	51.33	39.20	63.22	62.88	66.04	65.77	37.23	38.50	34.98	29.49
fastText	83.99	83.94	58.00	56.48	66.99	66.92	69.17	69.09	38.98	38.32	40.10	34.92
BERT	91.13	90.90	64.67	62.75	78.46	78.26	88.44	88.51	62.77	62.72	40.26	34.92
RoBERTa	**95.07**	**95.11**	67.67	66.07	80.56	80.56	88.75	88.76	66.86	67.53	41.18	36.54
MentalBERT	94.58	94.62	64.67	62.63	80.28	80.04	88.65	88.61	67.45	67.34	45.69	39.73
MentalRoBERTa	94.33	94.23	**70.33**	**69.71**	**81.82**	**81.76**	**88.96**	**89.01**	**68.61**	**68.44**	**50.48**	**47.62**
Zero-shot LLM-based Methods												
LLaMA-7B$_{ZS}$	63.55	58.91	57.0	56.26	54.83	53.51	23.04	25.55	10.53	11.04	13.92	16.34
LLaMA-13B$_{ZS}$	67.24	54.07	50.0	39.29	47.83	36.28	23.04	25.27	12.57	13.2	13.12	14.64
InstructGPT-3$_{ZS}$	58.87	58.66	50.33	49.86	50.07	49.88	27.60	26.27	12.70	9.36	10.70	12.23
ChatGPT$_{ZS}$	82.76	82.41	60.33	56.31	72.72	71.79	39.79	33.30	55.91	54.05	32.43	33.85
Emotion-enhanced CoT LLM-based Methods												
ChatGPT$_{V}$	79.51	78.01	59.20	56.34	74.23	73.99	40.04	33.38	52.49	50.29	28.48	29.00
ChatGPT$_{N_sen}$	80.00	78.86	58.19	55.50	70.87	70.21	39.00	32.02	52.92	51.38	26.88	27.22
ChatGPT$_{N_emo}$	79.51	78.41	58.19	53.87	73.25	73.08	39.00	32.25	54.82	52.57	35.20	35.11
ChatGPT$_{CoT}$	82.72	82.9	56.19	50.47	70.97	70.87	37.66	32.89	55.18	52.92	39.19	38.76
ChatGPT$_{CoT_emo}$	83.17	83.10	61.41	58.24	75.07	74.83	34.76	27.71	58.31	56.68	43.11	42.29
ChatGPT$_{CoT_emo_FS}$	**85.73**	**84.22**	**63.93**	**61.63**	**77.80**	**75.38**	**49.03**	**43.95**	**66.05**	**63.56**	**48.75**	**45.99**

Рис. 2: Результаты метрик

Результаты

- **Производительность моделей** (Таблица 1):
  - ChatGPT превосходит другие LLMs (F1: 82.41% для DR, 56.31% для CLPsych15), но уступает специализированным моделям (MentalRoBERTa: F1 89.01% для T-SID).
  - Эмоционально усиленные CoT-подсказки улучшают результаты (напр., +16.24% F1 для T-SID).
- **Объяснимость**:
  - ChatGPT генерирует объяснения, близкие к человеческим (средний балл 2.5/3.0 по всем критериям).
  - BART-Score демонстрирует наивысшую корреляцию с человеческой оценкой (Pearson: 0.590 для беглости).
- **Ограничения**:
  - Нестабильность предсказаний при изменении формулировок подсказок (напр., F1 варьируется от 47.55% до 71.79% для Dreaddit).
  - Ошибки в длинных контекстах из-за пропуска ключевой информации.

Рекомендации и выводы

- Использование few-shot обучения снижает вариативность предсказаний (напр., дисперсия F1 уменьшается с 89.29 до 31.93 для Dreaddit).
- Необходимость разработки специализированных метрик для автоматической оценки объяснений.
- Перспективы: тонкая настройка LLMs на данных, связанных с психическим здоровьем, и интеграция мультиязычных корпусов.

From benchmark to bedside: transfer learning from social media to patient-provider text messages for suicide risk prediction

Авторы: Hannah A. Burkhardt, Xiruo Ding, Amanda Kerbrat, Katherine Anne Comtois, Trevor Cohen

Аннотация: Исследование демонстрирует эффективность трансферного обучения из данных социальных сетей (Reddit) для улучшения прогнозирования суицидального риска в клинических текстовых сообщениях (программа Caring Contacts). Многоэтапное трансферное обучение с использованием предобученной модели PHS-BERT повысило F1-меру с 0.734 до 0.797. Разработана метрика клинической полезности (ATRIUM), оценивающая сокращение времени ответа на срочные сообщения. Результаты показывают, что автоматическая триажа сокращает задержки ответа на 15.9 минут для срочных случаев, подтверждая практическую ценность подхода.

Наборы данных

- **Социальные медиа (Reddit)**: 1105 постов от 621 пользователя (Shing et al., 2018). Аннотированы экспертами по уровню суицидального риска.
- **Клинические данные (Caring Contacts)**: 1229 сообщений от 221 участника рандомизированного испытания. Аннотированы по уровню срочности (0–3).
- **Ключевые различия**: длина сообщений (222 vs. 9 слов), анонимность, целевая аудитория.

Модели

- **Bag-of-Words + Логистическая регрессия**: Базовый подход с лемматизацией и стоп-словами.
- **PHS-BERT**: Предобученная модель для задач общественного здоровья.

- **Многоэтапное трансферное обучение**: Дополнительная предобучка на Reddit с адаптацией скорости обучения (Howard & Ruder, 2018).

Метрики оценки

- **Традиционные**: F1, Precision, Recall, AUC-ROC.
- **Клиническая полезность (ATRIUM)**:

$$\text{ATRIUM} = \left(\frac{a}{U}\right) T_{\text{urgent}} + \left(1 - \frac{a}{U}\right) T_{\text{nonurgent}} \tag{2}$$

где a — верно идентифицированные срочные сообщения, U — общее число срочных сообщений.

Результаты

Таблица 2: Сравнение моделей (медианные значения F1)

Модель	F1
Логистическая регрессия	0.585
PHS-BERT (без трансфера)	0.734
PHS-BERT + трансфер	0.797

Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data

Авторы: Сюйхай Сюй, Биншэн Яо, Юаньчжэ Дун, Саадия Габриэль, Хун Ю, Джеймс Хендлер, Марзие Гасеми, Анинд К. Дей, Дакуо Ван

Аннотация: В работе представлено комплексное исследование возможностей больших языковых моделей (LLM) для задач прогнозирования психического здоровья на основе текстовых данных из социальных сетей. Эксперименты с моделями Alpaca, FLAN-T5, GPT-3.5 и GPT-4 показали, что дообучение с инструкциями значительно улучшает производительность. Лучшие модели Mental-Alpaca и Mental-FLAN-T5 превзошли GPT-3.5 на 10.9% и GPT-4 на 4.8% по сбалансированной точности. Обсуждаются этические риски и ограничения, включая культурные и гендерные смещения.

Наборы данных

- **Dreaddit**: 2929 пользователей Reddit с аннотацией стресса (Turcan et al., 2019).

- **DepSeverity**: 4 уровня депрессии на основе DSM-5.
- **CSSRS-Suicide**: 500 пользователей с оценкой суицидального риска по шкале C-SSRS.
- Внешние наборы: Twitter (Twt-60Users) и SMS (SAD) для валидации.

Методы

- **Zero-shot prompting**: Шаблоны с контекстным и тематическим усилением.
- **Few-shot prompting**: Добавление примеров в промпты.
- **Instruction Finetuning**: Мультидатасетное дообучение Alpaca и FLAN-T5.

Метрики

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{3}$$

Результаты

Таблица 3: Сравнение моделей по сбалансированной точности

Модель	Dreaddit (Стресс)	DepSeverity (Депрессия)	CSSRS (Суицид)
GPT-3.5	0.688	0.653	0.617
GPT-4	0.725	0.719	0.760
Mental-Alpaca	**0.816**	**0.775**	0.730
Mental-FLAN-T5	0.802	0.759	**0.868**

Ключевые выводы

- Дообучение с инструкциями повышает точность на 14.7-23.4%.
- Модели на диалогах (Alpaca) лучше адаптируются, чем task-solving (FLAN-T5).
- GPT-4 демонстрирует продвинутые reasoning-способности, но ограничен в deployability.
- Выявлены этические риски: культурные стереотипы и генерация вредных советов.

Метод интеллектуального анализа текстовой информации для психиатрической диагностики

Авторы: Виктор Андреевич Петраевский, Алла Григорьевна Кравец

Аннотация: Автоматизированная система обнаружения депрессии на основе текстовых данных пациентов использует рекуррентные нейронные сети (LSTM) и векторизацию GloVe. Метод демонстрирует точность 93% на общедоступных англоязычных наборах данных, исключая искажение информации и обеспечивая объективность анализа. Уникальность подхода — сохранение контекстной семантики текста. Перспективы включают адаптацию модели для русского языка и улучшение контекстного анализа.

Наборы данных

- **Suicide and Depression Detection** (Kaggle): 11 000 постов Reddit (2008–2021), размеченных как «суицидальные» и «не суицидальные».
- **The Depression Dataset**: данные о связи депрессии с генетикой и образом жизни.
- Отсутствие русскоязычных датасетов: использованы англоязычные из-за их объёма и сбалансированности.

Метод

Этапы предобработки:

- Удаление пунктуации, цифр, эмодзи, стоп-слов.
- Лемматизация (NLTK) вместо стемминга для точности.
- Векторизация текста: сравнение GloVe (93% точности) и Word2vec (91% точности).

Архитектура модели:

- Слой Embedding с предобученными векторами GloVe.
- Слой LSTM (128 нейронов, dropout=0.2).
- Выходной слой Dense с сигмоидной активацией.

Таблица 4: Точность модели с разными методами векторизации

Метод	Точность
Word2vec	0.91
GloVe	0.93

Эксперименты

- Точность на тестовых данных: 0.9284.
- Функция потерь: сходимость после 4 эпох.

Результаты тестирования

Примеры классификации:

- *«I commit suicide»* — 92.54% (депрессия).
- *«The sun's warm embrace...»* — 1.66% (нет депрессии).
- Контекстно-сложные тексты (напр., метафоры) корректно распознаются.

Заключение

Метод показал высокую эффективность в автоматической диагностике депрессии. Планируется:

- Интеграция мультиязычных моделей.
- Использование трансформеров (BERT) для улучшения контекстного анализа.

Минусы статьи

- **Языковая ограниченность**: Модель обучена на англоязычных данных, что снижает применимость в русскоязычной среде.
- **Культурные различия**: Лингвистические маркеры депрессии могут варьироваться между культурами.
- **Переобучение**: Высокая точность на тестовых данных (92.84%) требует проверки на внешних наборах.
- **Отсутствие мультимодальности**: Учёт только текстовых данных, игнорирование интонации или визуальных маркеров.

Модели, которых нет в статьях, но они могут быть полезны

Я поискал в интернете модели, подходящие по теме, вот некоторые из них

Модели

PsychBERT

- **Описание**: Модель, дообученная на данных, связанных с психическим здоровьем.
- **Ссылка**: https://huggingface.co/mnaylor/psychbert-finetuned-mentalhealth

EmotionRoBERTa

- **Описание**: Модель, обученная на данных Twitter для анализа эмоций. Может быть полезна для выявления эмоциональных маркеров в текстах.
- **Ссылка**: https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion

BioClinicalBERT

- **Описание**: Модель, обученная на медицинских данных, включая тексты, связанные с побочными эффектами лекарств.
- **Ссылка**: https://huggingface.co/anindabitm/sagemaker-BioclinicalBERT-ADR

DistilBERT Mental

- **Описание**: Облегченная версия BERT, дообученная на данных, связанных с психическим здоровьем.
- **Ссылка**: https://huggingface.co/AventIQ-AI/distilbert-mental-health-prediction

llama-3.1-70B-finetuned-sentiment-analysis-for-mental-health

- **Описание**: Крупная модель, дообученная для анализа настроений в контексте психического здоровья. На ноуте не получится запустить, но при больших вычислительных мощностях выйдет
- **Ссылка**: https://huggingface.co/AhmedSSoliman/llama-3.1-70B-finetuned-sentiment-analysis
tree/main

mental-longformer-base-4096

- **Описание**: Модель Longformer, дообученная на данных, связанных с психическим здоровьем. Подходит для анализа длинных текстов.
- **Ссылка**: https://huggingface.co/AIMH/mental-longformer-base-4096

Еще один датасет большей величины, который можно использовать для LLM

Reddit Mental Health Posts

- **Описание**: Датасет, содержащий посты с Reddit, связанные с психическим здоровьем.
- **Ссылка**: https://huggingface.co/datasets/solomonk/reddit_mental_health_posts