CS447 Literature Review: Emotion or Mental State Recognition in Utterances in Texts using Natural Language Processing Methods

Shu Kit Tse
tse1@illinois.edu

December 9, 2024

Abstract

In this literature review, we read and summarize 4 papers in relation to our topic on emotion recognition or mental state/health recognition/detection, in particular to utterances in texts, as this is most common modality. The significance of improving emotion recognition or mental state/health recognition/detection is highly apparent in the mental health domain, while emotion recognition itself could also be applied broadly. After having reviewed 2 papers on LLMs in the mental health domain, and 2 papers that are more closely related to emotion recognition using NLP methods (mentioning each paper's contributions and relevance to our topic), we briefly discuss and conclude by mentioning the possibilities of using all of these methods altogether or integrating more methods or concepts into our NLP systems.

1 Introduction, Topic and Motivation

Utterances eg. conversational texts, comments found on various online platforms such as social media and chat apps while often lacking non-verbal cues (such as tone of voice, facial expression, body language) still likely convey sufficiently substantial information in a user's emotional and mental state, which may provide additional meaning to the user's utterances (besides the literal texts themselves, contexts involved and implications by a user), and where oftentimes other people may not necessarily be able to accurately decipher as well i.e. unable to "read between the lines". Instead of expanding the scope of this review to include utterances of all modalities eg. audio and video (where we would be able to obtain more information to improve emotion or mental state recognition), the topic is kept to utterances in texts as such utterances are the most prevalent in everyday use (where it is obvious that as consumers we more often generate texts eg. texts in chat apps, than we do audio or video eg. voice messages, social media shorts), hence work built upon emotion recognition, and mental state or mental health condition recognition for such a modality could potentially be applied more broadly.

The ability to accurately recognize a user's emotional and mental state based on their utterances (whether in an active ongoing online conversation, or in a user's social media posts etc.), and likely other contextual or individual profile information, would have numerous practical implications, many of which would hopefully be for good. A prime example would be to detect individuals who may be at risk of certain mental health conditions (eg. major depression, suicide ideation) that may require intervention (eg. by identifying and reaching out to at-risk individuals based on their social

media posts), or more generally just to check if someone may be mentally or emotionally sound. Another might be to assist online human therapists, or even an AI therapist (eg. a large language model (LLM) chat app), to conduct sessions more effectively and being able to potentially better empathize as needed.

While eventually accomplishing this in a marketed application may bring about ethical concerns also in particular to user privacy, this literature review does not attempt to partake in such debate or conversation. Here we aim to explore the work of 4 research papers that either advance work done in recognizing or detecting users' emotions or mental states (including mental health conditions) in utterances or more directly in creating research-oriented mental health applications such as a pretrained language model (PLM) which could likely perform such emotion recognition as well. We thus want to understand what are some relatively recent models and techniques in natural language processing (NLP), including PLMs (and thus by definition including LLMs), that advance such work and what are they currently capable of.

In doing so, here are some questions at large that we look to explore, which may or may not be well answered by the end of this literature review. What types of mental health conditions or emotional states are these models and techniques capable of detecting? How would emotions or mental states be categorized formally in context of emotion or mental health recognition or analysis by a system? What would such recognition or analysis typically be based on? What are the methods used to perform recognition? How effective are they and how are they typically evaluated? Are LLMs in particular effective in such recognition or analysis? What datasets are used in these research efforts, and how reliable (eg. representation bias) are the results based on such datasets used? What other related psychological capabilities are reasonably capable by such models, systems or NLP methods? What else could reasonably be considered in future efforts?

As a side note, while I personally refer to the term mental state, more often than not, it appears that it is merely easier to find work related to the term and concept of mental health. The term mental state, as I understand it, would include a person's emotional, cognitive (eg. understanding, attention, memory) and physiological (eg. fatigue, momentary stress) states as well as for example intentions. As mental health conditions for example would form part of a person's persistent mental state, I personally think covering work done in mental health while keeping the term mental state in this review to indicate a more comprehensive perspective remains appropriate.

2 Papers Selected

1. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare (Ji et al., 2022)

2. Towards Interpretable Mental Health Analysis with Large Language Models (Yang et al., 2023)

3. Language and Mental Health: Measures of Emotion Dynamics from Text as Linguistic Biosocial Markers (Teodorescu et al., 2023)

4. Emotion Recognition in Conversation via Dynamic Personality (Wang et al., 2024)

As I searched and browsed papers on ACL Anthology related to the topic, I gradually came across papers whose ideas and methods look to be more diverse than what I initially looked for (which were in LLMs, and emotion recognition and mental health detection). Nevertheless, as

PLMs are undoubtedly powerful methods for various NLP tasks, I needed to learn for myself and see what might or could currently be considered state-of-the-art (at least based on what I can find) in research related to NLP in emotion recognition and mental health detection.

I soon came across topics in emotion dynamics (specifically in utterance emotion dynamics) and emotion recognition in conversation (ERC), which were topics directly relevant to what I had in mind, and thus it would have been remiss of me not to include these papers (3 and 4) in this review and later discussion on existing methods as both involve interdisciplinary methods in psychology alongside NLP.

This set of 4 papers was selected as such to represent the different kinds of research work done that fall under this literature review's topic. The first paper is on training a domain-specific PLM, which is now used as a benchmark/baseline in mental health related NLP research. The second more directly compares domain-specific and supervised methods against LLMs on mental health detection tasks alongside other tasks, showcasing the capabilities of LLMs in the mental health domain. The third paper illustrates how we could incorporate concepts in psychology, in particular to emotion dynamics in utterances, used to prove and establish a relationship between emotion dynamics and mental health conditions, upon which future work could be built. Lastly, the fourth paper more directly integrates the concept of dynamic personality from psychology into a novel framework to improve emotion recognition prediction in conversations, showcasing how something as abstract as personality could be integrated into a model to a certain extent; again which proves there is more we could do along these lines and in these related domains. All in all, this set of 4 papers showcases different NLP methods in relation to emotion recognition and/or mental state/health detection/recognition in utterances in texts, where it would be interesting to imagine what a combination of some of these methods and beyond could possibly produce.

3 Note on Background

This literature review assumes the reader to have a basic knowledge of what LLMs are, what prompting is, as well as general basic knowledge about machine learning. Paper 3 in particular would likely require some explanation on what the topic of emotion dynamics is, and a couple of brief paragraphs has been included within the review itself to facilitate the sharing of such background knowledge, and hence the omission of a background section here altogether. Certain concepts in psychology such as the use of the Big Five Model in personality in paper 4 could likely either be considered general knowledge or could also easily be understood from a brief online search.

4 Paper 1 Review - MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare

In 2022 (this paper (Ji et al., 2022) was published in June 2022), while there were pretrained language models (PLMs) in the medical domain (BioBERT) and in clinical notes (ClinicalBERT), there was not any in the mental healthcare domain. Mental health, as we all know, is a critical issue in society, where with early detection, mental disorders can be helped with effective social interventions, which led the authors to create MentalBERT and MentalRoBERTa, two PLMs in the mental healthcare domain meant for use by the research community. MentalBERT and MentalRoBERTa were evaluated using several mental disorder detection benchmarks where improvement was shown for mental health detection tasks against then existing baselines.

MentalBERT and MentalRoBERTa, as obvious from their names, were trained from BERT and RoBERTa using Huggingface's Transformer framework (Wolf et al., 2020) and a domain-adaptive pretraining scheme (Gururangan et al., 2020) that continues pretraining in the downstream domain of mental health. The pretraining corpus consists of posts in selected mental health-related subreddits (Reddit communities with specific topics of interest) including "r/depression", "r/SuicideWatch", "r/Anxiety", "r/offmychest", "r/bipolar", "r/mentalillness", and "r/mentalhealth", totaling at 13,671,785 sentences.

Category	Platform	Dataset	train	validation	test
Assorted	Reddit	SWMH (Ji et al., 2022)	34,823	8,706	10,883
Depression	Reddit	eRisk18 T1 (Losada and Crestani, 2016)	1,533	658	619
Depression	Reddit	Depression Reddit (Pirina and C¸ oltekin, 2018)	1,004	431	406
Depression	Reddit	CLPsych15 (Coppersmith et al., 2015)	457	197	300
Stress	Reddit	Dreaddit (Turcan and McKeown, 2019)	2,270	568	715
Suicide	Reddit	UMD (Shing et al., 2018)	993	249	490
Suicide	Twitter	T-SID (Ji et al., 2022)	3,072	768	960
Stress	SMS-like	SAD (Mauriello et al., 2021)	5,548	617	685

Table 1: Summary of datasets used in evaluating MentalBERT and MentalRoBERTa (Paper 1 Review)

The pretrained MentalBert and MentalRoBERTa were then further fine-tuned in binary mental disorder detection and multi-class mental disorder classification, and evaluated against several datasets in depression, suicidal ideation, and other mental disorders (eg. stress, anxiety, bipolar). An example of a dataset used is the CLPsych 2015 Shared Task dataset (Coppersmith et al., 2015) containing posts from users with depression on Twitter, where the train partition used consists of 327 depression users, and test data 150 depression users. See table 1 for a summary of the datasets used, directly obtained from and the same as shown in the paper itself.

Model	eRisk T1		CLPsych		Depression Reddit	
	Rec.	F1	Rec.	F1	Rec.	F1
BERT	88.53	88.54	64.67	62.75	91.13	90.90
RoBERTa	92.25	92.25	67.67	66.07	**95.07**	**95.11**
BioBERT	79.16	78.86	65.67	65.50	91.13	90.98
ClinicalBERT	76.25	75.41	65.67	65.30	89.41	89.03
MentalBERT	86.27	86.20	64.67	62.63	94.58	94.62
MentalRoBERTa	**93.38**	**93.38**	**70.33**	**69.71**	94.33	94.23

Table 2: Results of depression detection. The bold text represents the best performance.

Tables 2 (depression datasets) and 3 (datasets for other disorders) are also directly from the paper, and show recall and F1 evaluation results for MentalBERT and MentalRoBERTa compared to various baselines. The F1 score metric is used as mental disorder detection is usually a task with class imbalance. As can be seen from the results summary tables, MentalRoBERTa performed best on eRisk T1 and CLPsych for the depression datasets, while RoBERTa still fared better on Depression Reddit. For the other datasets, MentalBERT's recall comes up on top for the UMD dataset, while RoBERTa did best on F1 for UMD and recall for SWMH. MentalRoBERTa scored

Model	UMD		T-SID		SWMH		SAD		Dreddit	
	Rec.	F1	Rec.	F1	Rec.	F1	Rec.	F1	Rec.	F1
BERT	61.63	58.01	88.44	88.51	69.78	70.46	62.77	62.72	78.46	78.26
RoBERTa	59.39	**60.26**	88.75	88.76	**70.89**	72.03	66.86	67.53	80.56	80.56
BioBERT	57.76	58.76	86.25	86.12	67.10	68.60	66.72	66.71	75.52	74.76
ClinicalBERT	58.78	58.74	85.31	85.39	67.05	68.16	62.34	61.25	76.36	76.25
MentalBERT	**64.08**	58.26	88.65	88.61	69.87	71.11	67.45	67.34	80.28	80.04
MentalRoBERTa	57.96	58.58	**88.96**	**89.01**	70.65	**72.16**	**68.61**	**68.44**	**81.82**	**81.76**

Table 3: Results of classifying other mental disorders including stress, anorexia, suicidal ideation. The bold text represents the best performance.

best for the rest of the datasets evaluated and remaining recall and F1 metrics.

This paper's key contribution was in training and releasing MentalBERT and MentalRoBERTa, two PLMs pretrained in the mental health domain, and making them publicly available for research and usage. The evaluation results further show that MentalRoBERTa in particular outperforms then existing PLM baselines, showcasing the potential of PLMs (and thus large language models (LLMs)) for mental health condition detection.

In my opinion, the significance of this paper and the work done was in pushing the boundaries of PLMs in the mental health domain (where we have already mentioned the domain's importance, and where to the authors' knowledge this was the first work to have trained domain-specific language models in this mental health), and providing access to such otherwise valuable resources (since PLMs are typically expensive to train) by making them publicly available.

In this review, we look to understand what are some existing NLP methods on emotion or mental state recognition based on user utterances, where this paper has done the latter. By achieving competitive results in particular to those of MentalRoBERTa's, this work first showcased PLMs' capabilities in mental health condition detection and second formed a benchmark upon which future work can be compared to, which we later also see used in paper 2.

5 Paper 2 Review - Towards Interpretable Mental Health Analysis with Large Language Models

This paper (Yang et al., 2023), published in December 2023, studied large language models' (LLMs) capabilities in performing mental health analysis (and emotional reasoning), alongside their ability to generate explanations for their predictions on 11 datasets across 5 tasks.

While the importance of using NLP methods and systems in the mental health domain is apparent so far in this review, the authors' purpose in this work was to resolve certain issues in helping to push forward a new LLM-based paradigm in mental health analysis, resolving various issues. One such issue was the general lack of explainability on mental health detection results, where cause/factor detection of mental health conditions were also generally ignored. Another issue was that previous works used simple prompts to detect mental health conditions, which ignored the usage of certain useful information such as emotional cues.

The four LLMs used in this research study were ChatGPT (gpt-3.5-turbo), InstructGPT-3, LLaMA-13B, and LLaMA-7B. In particular, this version of ChatGPT was trained based on the 175-billion-parameter version of InstructGPT, continually optimized through reinforcement learning from human feedback. Similar to paper 1, the datasets were meant for mental health condition detection

You will be presented with a post. Consider the emotions expressed in this post to identify whether the poster suffers from [condition]. Only return Yes or No, then explain your reasoning step by step. Here are N examples:
Post: [example 1]
Response: [response 1]
...
Post: [example N]
Response: [response N]

Post: [Post]
Response:

Figure 1: Few-shot emotion-enhanced prompting template, as originally shown in paper 2

eg. depression, suicide, stress, and were known datasets in the mental health domain i.e. DepressionReddit, CLPsych15, Dreaddit, T-SID, SAD, and CAMS.

The authors explored the effects of different prompting strategies, namely in 1) zero-shot prompting, 2) emotion-enhanced Chain-of-Thought (CoT) prompting (where the prompt also asked the LLMs to generate step-by-step explanations), 3) few-shot emotion-enhanced prompting and 4) distantly supervised emotion-enhanced prompting (involving the usage of VADER and NRC EmoLex sentiments, and NRC EmoLex emotions).

Few-shot emotion-enhanced prompting (with CoT) involved having domain experts (Ph.D. students in quantitative psychology) writing one response example for each label class within a test set (each response consisting a prediction and an explanation behind the decision) and using this to enable in-context learning; see Figure 1 for an example of the template. Distantly supervised emotion-enhanced prompting involved using lexicons to assign a sentiment or emotion score to each post and labeling the post based on such a score, then adding these sentiment/emotion labels to the proper positions of the zero-shot prompt (hence providing emotion-related information in the prompt).

The LLMs and prompting methods were evaluated against baseline models on classification tasks using the aforementioned datasets. The baseline models measured against can be seen in table 4, where we note that PLMs such as BERT and RoBERTa, as well as domain-specific ones i.e. MentalBERT and MentalRoBERTa were also used for comparison.

Here we note that while ChatGPT outperforms other LLMs in zero-shot prompting, various emotion-enhanced prompting methods (using ChatGPT) were actually less effective compared to zero-shot (a reason , where the only real improvement came from using emotion-enhanced CoT prompting with ChatGPT, with the few-shot method coming out further on top. As the authors mentioned, this thus proved that emotion-enhanced CoT prompting is effective for mental health analysis.

We also however note that although ChatGPT outperformed the other three LLMs, it still underperforms fine-tuning-based methods where MentalRoBERTa and RoBERTa actually performed best across the datasets (as can be seen from the bold numbers in table 4), which left room for improvement in terms of further exploring then LLMs' mental health analysis ability.

Model	DR		CLPsych15		Dreddit		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	66.09	66.02	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	80.04	80.03	88.65	88.61	67.45	67.34	45.69	39.73
MentalRoBERTa	94.33	94.23	**70.33**	**69.71**	**81.82**	**81.82**	**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	36.22	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	70.07	52.92	39.00	32.02	51.92	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**

Table 4: Test results on the mental health analysis tasks, as directly obtained from paper 2 (Yang et al., 2023). ChatGPT$_{V}$, ChatGPT$_{N_sen}$, and ChatGPT$_{N_emo}$ denote the emotion-enhanced prompting methods with VADER sentiments, NRC EmoLex sentiments, and NRC EmoLex emotions. ChatGPT$_{CoT}$ and ChatGPT$_{CoT_emo}$ denote the zero-shot and emotion-enhanced CoT methods on the corresponding task. ChatGPT$_{CoT_emo_FS}$ combines expert-written few-shot examples in the emotion-enhanced prompt. The results of baseline methods are referenced from (Ji et al., 2022).

Human evaluation was also done to assess the quality of generated explanations by ChatGPT and InstructGPT-3 for the same emotion-enhanced prompts. Four key aspects in fluency (coherence and readability of explanation), reliability (trustworthiness of explanations to support predictions), completeness (how well explanations cover relevant aspects of original post), and overall (general effectiveness of explanation) were assessed, with ChatGPT outperforming InstructGPT-3, and indicating ChatGPT's potential to generate approaching-human explanations.

As cause/factor detection is not really relevant to our topic in this review, we exclude our discussion on this part of paper 2's work. In this work, ChatGPT also showed limitations in unstable predictions caused by excessive sensitivity to minor alterations in prompts eg. using different adjectives to indicate degree of seriousness of mental health condition, and inaccurate reasoning eg. ignorance of relevant information in long posts and unreliable reasoning process.

The authors further suggested and indicated future work on domain-specific fine-tuning LLMs to improve inaccurate reasoning problems, and concluded that while mental health analysis may still be challenging for LLMs, using proper prompt engineering with emotional information could improve results, and that ChatGPT was capable of generating human-level explanations for its decisions.

Based on my understanding, this paper's contributions were in comprehensively evaluating LLMs on mental health detection tasks, also with an emphasis on using different prompt methods as well as on whether LLMs were capable of generating accurate quality explanations for their decisions. In doing so, the authors also created a novel dataset of 163 human-assessed explanations used in human evaluation of ChatGPT's and InstructGPT-3's prediction explanations.

In my opinion, this paper is relevant to this literature review's topic on emotion recognition or mental state recognition in user utterances in the sense that LLMs were shown to be capable of detecting mental health conditions (although then worse off than MentalRoBERTa and RoBERTa) and could be improved with the right prompting approaches. We could thus easily imagine instead of classifying i.e. detecting mental health condtions, we could have LLMs detect different emotions instead, where similar prompting methods as well as prediction explanations could be utilized. Furthermore, as shown by comparison of LLMs against baseline models, we know that although ChatGPT during this time was largely outperformed by a domain-specific fine-tuned model i.e. MentalRoBERTa, we could similarly imagine potentially having a domain-specific fine-tuned LLM in both emotion recognition and mental state (including mental health) recognition/detection, also further specific to data involving utterances in texts.

6 Paper 3 Review - Language and Mental Health: Measures of Emotion Dynamics from Text as Linguistic Biosocial Markers

The authors in paper 3 (Teodorescu et al., 2023) studied the relationship between tweet emotion dynamics and mental health disorders, through measures of utterance emotion dynamics (UED) from tweets of users who self-reported as having a mental health diagnosis, where this ultimately shows how emotion dynamics in utterances can serve as biosocial markers for mental disorders and help in such detection and management.

The paper was motivated by several aspects, where for starters the patterns of a person's emotional change over time (i.e. emotional dynamics) are actually indicators of one's mental health and well-being as shown by research in psychopathology. Secondly, as language is 1) inherently social, 2) where aspects of language that can be measured have been shown to act as biomarkers (i.e. associated with a disease outcome or biology in general), 3) where social factors eg. parental socioeconomic status, neighbourhood, influence speech markers, and 4) how emotions expressed in text have been shown to correlate with mental health diagnosis, language can thus be considered as a biosocial marker for health i.e. influenced by both social and biological factors. Thus, interestingly, we should therefore be able to find out and show that measures of emotion dynamics from one's use of language eg. utterances could act as biosocial markers for mental health (which once again as we all know is an important domain).

More specifically, the authors wanted to see how UED metrics in 1) average emotional state, 2) emotional variability, 3) rate at which emotions reach peak emotional state (i.e. rise rate) and 4) rate at which emotions recover from peak emotional state back to steady state (i.e. recovery rate), differ between a control group and groups of Twitter users who self-disclosed to having one of seven mental health conditions (MHC) i.e. ADHD (attention-deficit/hyperactivity disorders), bipolar, depression, MDD (major depressive disorder), OCD (obsessive-compulsive disorder), PPD (post-partum depression), PTSD (post-traumatic stress disorder). Each of the UED metrics (again, compared between a control group and each MHC group) are further broken down into three dimensions of emotion i.e. valence, arousal and dominance.

Let us first briefly go through emotion dynamics as the authors have done so, based on my understanding. To start, average emotional state refers to the baseline or average emotional state a person is at, where for example individuals with mental illnesses tend to have more negative emotional baselines. Emotion variability refers to the degree of fluctuations or changes in emotional states over time, where having higher emotion variability i.e. where a person frequently shifts between different emotional states (eg. from happy to sad or angry) has been linked to lower psychological

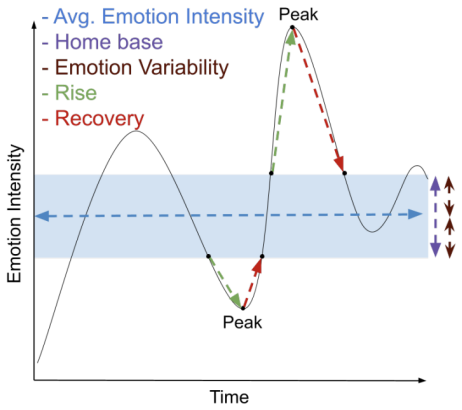

Figure 2: Diagram visualizing utterance emotion dynamics metrics

well-being. Emotion reactivity refers to the intensity and speed of emotional responses to positive or negative events, where it has been shown that individuals with psychopathologies tend to take longer to recover from differing emotional states than healthy individuals i.e. difficulty moving between emotional states is linked to lower psychological well-being. This is also where emotion reactivity is related to rise rate and recovery rate mentioned above. Research in emotion dynamics suggests that each of these aspects may vary by a person's mental health, and across different mental illnesses.

The utterance emotion dynamics (UED) framework uses metrics inspired by psychology to quantify patterns of emotion change in text, allowing researchers to analyze emotion dynamics in people's utterances (which would reasonably reflect each individual's thought process). See figure 2 for a visual diagram. In this paper in particular, the core dimensions of emotion in valence (pleasure-displeasure or positive-negative), arousal (active-sluggish), and dominance (in control-out of control, powerful-weak) are also explored as part of the relationship between UED and mental health conditions.

The main dataset used (STMHD) in this work comprises of tweets from 27003 users self-reported as having a MHC on Twitter, with four years of tweets collected for each user (two years before self-reporting a MHC and two years after), between January 2017 and May 2021. A control group's tweets was randomly sampled from the period. A secondary dataset from Reddit (eRisk 2018) consists of users who self-disclosed as having depression on Reddit and a control group. More information on the dataset can be found in table 5.

Each user's tweets are ordered by timestamp, where the Emotion Dynamics toolkit (Vishnubhotla and Mohammad, 2022) (Hipson and Mohammad, 2021) was used to compute UED metrics (average emotion, emotion variability, rise rate, recovery rate). The authors then performed analyses for valence, arousal and dominance. Post hoc analyses for pairwise comparisons for each metric across these three dimensions of emotions were then performed to obtain results in figure 3.

We can generally note that most MHC groups compared to the control group had lower valences, arousal, and dominance for average emotion. Valence, arousal and dominance for emotion variability, rise rate and recovery were either neutral or higher for MHC groups compared to the control group. With the exception of how higher recovery rates (proxy for emotion regulation) ought to typically indicate higher psychological well-being (which is the case here based on results

Dataset	Group	#People	Avg. #Posts/User
Twitter	MHC	10,069	2,177.4
	ADHD	3,866	2,122.2
	Bipolar	721	3,193.3
	Depression	3,017	2,084.0
	MDD	133	2,402.9
	OCD	605	1,822.9
	PPD	105	1,671.4
	PTSD	1,622	1,944.9
	Control	4,097	1,613.6
Reddit	Depression	106	233.79
	Control	749	359.74

Table 5: The number of users in each mental health condition and the number of tweets per user in the pre-processed version of the Twitter-STMH and Reddit eRisk datasets used in paper 3 (Teodorescu et al., 2023).

		Average Emotion			Emotion Variability			Rise Rate			Recovery Rate		
Dataset	**MHC–Control**	V	A	D	V	A	D	V	A	D	V	A	D
Twitter-STMHD	ADHD–control	↓	↓	↓	↑	↑	↑	–	–	↑	–	↑	↑
	Bipolar–control	–	↓	↓	↑	↑	↑	–	–	–	↑	–	–
	MDD–control	↓	–	↓	↑	↑	↑	↑	–	–	↑	↑	↑
	OCD–control	–	↓	↓	↑	↑	↑	–	–	↑	–	↑	↑
	PPD–control	–	↓	↓	–	↑	↑	–	–	–	–	–	–
	PTSD–control	↓	–	↓	↑	↑	↑	↑	↑	–	↑	↑	↑
	Depression–control	–	↓	↓	↑	↑	↑	↑	–	↑	↑	↑	↑
Reddit eRisk	Depression–control	–	–	↓	↑	–	↑	–	–	–	–	–	–

Figure 3: Results of the difference in UED metrics between each MHC group and the control. A significant difference is indicated by an arrow; arrow direction indicates the direction of the difference.

between MHC groups and control group), the other three UED metrics (average emotion, emotion variability, rise rate) were aligned to results in psychology. The difference here might possibly be due to differences in mediums and collection process, or also just the fact that this is based on utterances rather than self-reported by individuals as is typically done in psychology. Biosocial aspects of UED metrics were also measured based on user popularity on Twitter (i.e. via average no. of likes on posts), and showed that the UED metrics had the same differences (direction wise) when comparing MHC groups to the control group.

The authors thus showed for the first time that emotion dynamics in individuals' utterances are significantly related to mental health conditions when compared to a control group, which provide important contextual information and indicators for mental health condition detection and management.

In my opinion, while the method on UED metrics with utterances in temporal order measuring 4 aspects of emotion dynamics (average emotion, emotion variability, rise rate, recovery rate) across

3 dimensions (valence, arousal, dominance) looks to be fantastic work that advances how we could perform emotional recognition and mental state/health recognition/detection, I am personally not entirely convinced by the mere use of average number of post likes on Twitter as a feature or indicator of the social aspects in using emotion dynamics in utterances as biosocial markers for mental health and well-being. While obtaining data that would be representative and indicate adequate social aspects (eg. sociodemographics) in users' language (i.e. utterances) might prove difficult, eventually being able to do so would make the authors' case more convincing.

In terms of relevance to this review's topic on emotion recognition and mental state/health recognition in utterances in text, I personally enjoyed understanding (reasonably so) what the UED metrics meant and represented, and found it interesting how we could use such interdisciplinary methods to measure and model psychological aspects of individuals via NLP methods. I would think that the concept of UED metrics being measured across the three core dimensions of emotion most certainly ought to form a part of any good approach in NLP systems looking to perform emotion recognition or mental state/health detection.

7 Paper 4 Review - Emotion Recognition in Conversation via Dynamic Personality

The authors of paper 4 (Wang et al., 2024) worked on improving emotion recognition in conversation (ERC) by introducing and integrating dynamic personality into a model (called ERC-DP in this paper), which accounted for past utterances from a person impacting slight shifts in this same person's dynamic personality throughout a conversation.

ERC aims to classify the emotion of each utterance within conversations, and in doing so handling emotional ambiguity across different speakers and contextual factors (this is almost a direct quote from paper 4). While static personality has been previously introduced as a deep speaker factor for emotion recognition, the dynamic variability aspect has very often been overlooked. This paper thus integrated dynamic personality of speakers into ERC, coming up with a novel model ERC-DP, and showed improved performance on three benchmark conversational datasets.

The Big Five Model of personality was used to represent personality states in ERC-DP, which comprises three components, namely 1) a personality recognition module, 2) a prompt design module, and 3) a fine-grained classification module. See figure 4 for an illustration of the architecture.

The personality recognition module consists of several components, the first was a prompt-based sentence encoder built upon BERT, which was fine-tuned using the Adam optimizer. The input into this BERT component is a prompt Prompt_P = The personality is [MASK] alongside the last and current utterances, which outputs a classification vector as a feature representation of dynamic personality. This then gets fed into a fully connected layer, then a ReLU function, followed by a Dropout layer to mitigate overfitting, and finally a MLP for classification. The entire module is trained using the personality recognition dataset Essays (Pennebaker and King, 1999) consisting 2468 anonymous essays tagged with personality traits, and outputs a 5-D personality score for each dimension of the Big Five Model with either 0 or 1 representing the dynamic personality from the current utterance. This then gets fed into the prompt design module.

The prompt design module first converts the 5-D personality score (representing personality dimensions openness, conscientiousness, extraversion, agreeableness, and neuroticism) into corresponding adjectives i.e. text. An example would be converting [1, 1, 0, 0, 1] into "This person is open, conscientious, not extraverted, not agreeable, and neurotic." This portion of the input is then

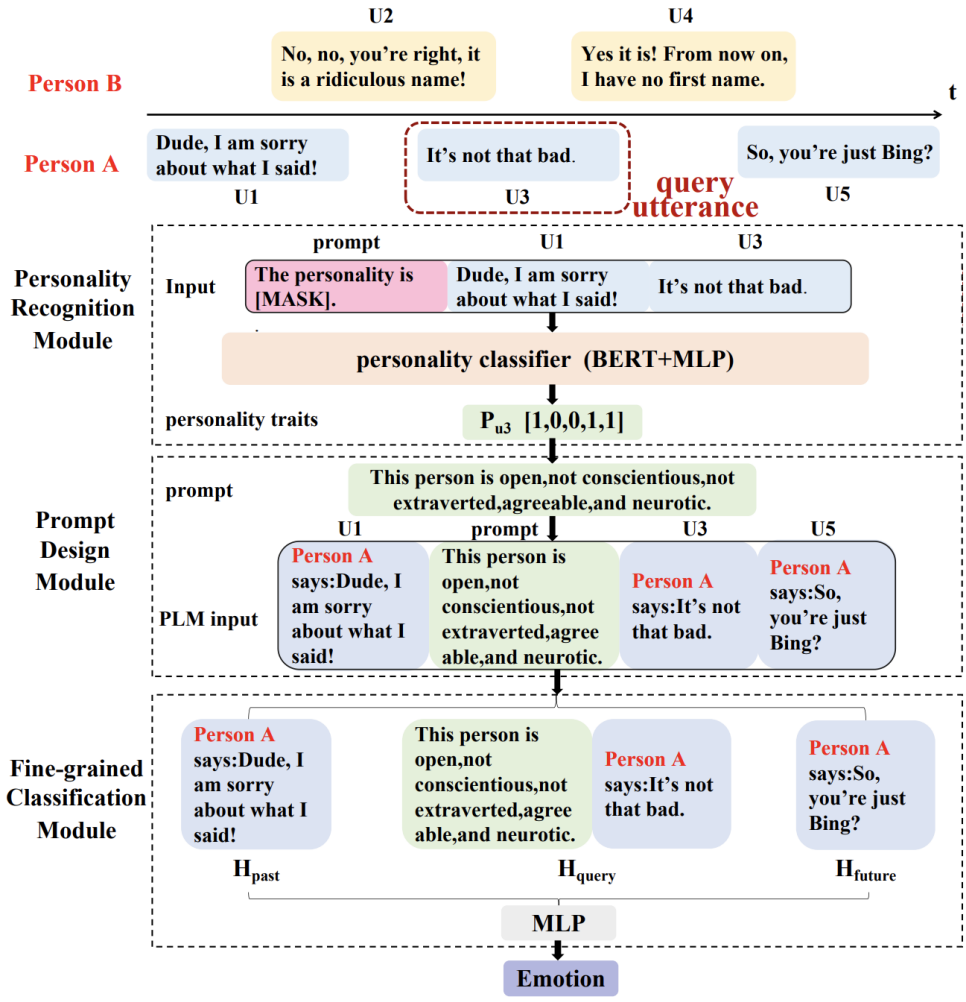

Figure 4: ERC-DP architecture

appended with as much context (speaker utterances) as possible via an algorithm specified in the paper, before being passed onto the next module.

The fine-grained classification module first consists of SimCSE (Gao et al., 2021) (with first eight encoder layers frozen, and applied Adam optimizer), where it obtains the last hidden vector $H_{\text{total}} \in R^{l*d}$ after passing the previous module's input into this first component, where l is the number of tokens. These l tokens are then split into 3 parts representing past features, query features and future features, which are then concatenated and passed into a classifier consisting of a fully connected layer, Tanh activation, Dropout layer, and a MLP to predict emotion for the current utterance using Focal Loss (Lin et al., 2017).

Three conversational emotion recognition datasets were used to train and evaluate the ERC-DP model, namely MELD, EmoryNLP, and IEMOCAP. See figure 5 for more details. MELD is a dataset based on Friends (TV show) comprising more than 1400 conversations and 13000 utterances labeled with one of seven emotions (anger, disgust, sadness, joy, surprise, fear, neutral). EmoryNLP

Dataset	Conversations			Utterances		
	Train	Val	Test	Train	Val	Test
IEMOCAP	120		31	5810		1623
MELD	1038	114	280	9989	1109	2610
EmoryNLP	713	99	85	9934	1344	1328

Figure 5: Details of conversational emotion recognition datasets used

is also based on Friends, and consists of 897 scenes and 12606 utterances, each labeled with one of seven emotions (neutral, joyful, peaceful, powerful, scared, mad, sad). IEMOCAP consists of two-person conversations among ten speakers, with five sessions total, each utterance labeled with one of six emotions (neutral, happiness, sadness, anger, frustration, excited).

Type	**Method**	**MELD**	**EmoryNLP**	**IEMOCAP**
Content Modeling methods				
	HiGRU (Jiao et al., 2019)	56.81	34.48	58.54
	DAG-ERC (Shen et al., 2021b)	63.65	39.02	68.03
Speaker Modeling methods				
	DialogRNN+RoBERTa (Majumder et al., 2019)	63.20	37.75	63.92
	COSMIC (Ghosal et al., 2020)	65.21	38.11	65.28
	DialogueCRN (Hu et al., 2021)	63.42	38.91	66.33
	SKAIG (Lee and Lee, 2022)	65.18	38.88	66.98
	DialogXL (Shen et al., 2021a)	62.41	34.73	66.20
	EmoBERTa (Kim and Vossen, 2021)	66.51	–	68.57
	COMPM (Lee and Lee, 2022)	66.52	38.93	69.46
	EmotionIC (Yingjian et al., 2023)	66.40	40.01	69.61
	SACL (Hu et al., 2023)	66.45	39.65	69.22
	BERT-ERC (Qin et al., 2023)	67.11	39.84	**71.70**
ERC-DP (paper 4's model)		**67.34**	**40.10**	69.64

Table 6: Comparison with baselines on three datasets

ERC-DP was compared to various baselines shown in table 6, where ERC-DP outperformed all existing baselines including state-of-the-art BERT-ERC (in ERC tasks) on MELD and EmoryNLP, but did not fare as well when it came to IEMOCAP. Weighted F1 score was used as the evaluation metric due to class imbalance in all three datasets. The lack of performance on IEMOCAP was likely attributed to potential inaccuracies in the dynamic personality predictions, stemming from two primary factors - length and nature of conversations in the dataset. The lengths of conversations for IEMOCAP were particularly long, affecting the length of the input texts for the personality module. The nature of the conversations on the other hand in IEMOCAP were much more diverse, as opposed to scripted TV show conversations in the other two datasets.

The authors also did an ablation study, removing the personality recognition module, and com-

Method	MELD	EmoryNLP	IEMOCAP
No Personality	66.70	39.10	68.14
Static Personality	67.08	39.47	69.02
Dynamic Personality	**67.34**	**40.10**	**69.64**

Figure 6: Ablation study on three datasets

paring the ablated version to ERC-DP (with dynamic personality) as well as a version with a static personality module, showcasing improvement in results using dynamic personality (see figure 6).

This paper thus contributed by coming up with and training a novel method ERC-DP which integrated dynamic personality into the ERC task, and showed improvements on the MELD and EmoryNLP datasets when compared against baselines including state-of-the-art BERT-ERC. The authors also showed via an ablation study that the dynamic personality module did in fact contribute to improving the overall model for ERC tasks.

The integration of dynamic personality into the ERC task, which is directly relevant to this review's topic, is most certainly interesting and impressive, and I think it would be good to dive deeper into why BERT-ERC still outperformed ERC-DP on IEMOCAP, potentially considering how to broaden the scope of potential ERC tasks to consider based on such findings. While it made sense that the baseline models being evaluated against are all built and trained for the ERC task, it might have been interesting or might perhaps yield certain insights if LLMs were used for comparison too. Although from paper 2, we could tell that domain-specific models already did outperform ChatGPT (gpt-3.5-turbo; in 2023), thus ERC task-specific models might very well likely do the same (nevertheless it would be nice to see).

Personally, I find the usage of a 0 or 1 5-D vector to represent the Big Five personality traits a bit too simplistic, as the personality construct is most likely much more nuanced than this. Even using continuous values i.e. a spectrum on each personality trait, as it is how it is likely done in psychology would potentially be more useful. On the flip side, it is also understandable that there are not any readily available datasets for such training, seeing as how the authors had to use a dataset from 1999 for the personality module. In any case, this was most certainly an interesting approach to integrate dynamic personality into the ERC task, where we could draw inspiration from this and translate many other social and psychological aspects of conversation and speakers involved into a future novel NLP framework to train and model.

8 Discussion and Conclusion

From the onset, we considered a set of 4 papers that were relevant to our topic on emotion recognition or mental state/health recognition/detection in utterances in texts, where each paper brings its own NLP methods in relation to our topic in this literature review.

To quickly recap, paper 1 (Ji et al., 2022) trained and made publicly available two PLMs MentalBERT and MentalRoBERTa in the mental health domain, showing improvements against then baselines in other related domains on most datasets evaluated.

Paper 2 (Yang et al., 2023) utilized various prompting methods across four LLMs, including

zero-shot, emotion-enhanced chain-of-thought (CoT), and distantly supervised (VADER and NRC EmoLex lexicons) emotion-enhanced prompts (both zero-shot and few-shot) and found that only emotion-enhanced CoT prompting (both zero-shot and few-shot) performed better than pure zero-shot prompting, proving that emotion-enhanced prompting does improve results. Of the four LLMs, ChatGPT (gpt-3.5-turbo) performed the best, but still underperformed when compared to baselines in MentalRoBERTa and RoBERTa, showing that domain-specific supervised methods are still better. The authors also showed ChatGPT's capability in explaining its prediction decisions at approaching human-level, and was much better than those of InstructGPT-3's.

Paper 3 (Teodorescu et al., 2023) showed that there are significant relationships between emotion dynamics in utterances and mental health conditions. This is also where UED metrics were measured in a highly structured manner i.e. via average emotion, emotion variability, rise rate, and recovery rate, and further for each across three core dimensions of emotions in valence, arousal and dominance (which likely require further reading to more deeply understand). This paper thus showcased the potential of using emotion dynamics in an individual's utterances to associate with their mental state (in this particular case, mental health status).

Paper 4 (Wang et al., 2024) integrated the prediction and usage of dynamic personality into the task of emotion recognition in conversations (ERC), showcasing a framework that outperformed existing baselines including state-of-the-art BERT-ERC on two datasets MELD and EmoryNLP, but underperformed compared to BERT-ERC in IEMOCAP, a dataset that is different from the other two.

While I have already briefly discussed what each paper's contributions were, and where appropriate my personal assessment on certain parts of each paper, as well as how each paper relates to this review's topic in emotion recognition or mental state/health recognition/detection in utterances in texts, in the remainder of this section I would like to combine some of these takeaways, making a personal conclusion of sorts from reading these papers and learning from this review.

My main takeaway from this literature review was learning what NLP methods were available in research in performing emotion recognition or mental state/health recognition/detection, and understanding how research is done in these related topics. Having learned how fine-tuned PLMs such as MentalRoBERTa and MentalBERT could outperform ChatGPT (gpt-3.5-turbo) in 2023 in mental health condition detection was not something I personally expected (given the differences in the number of model parameters), where I picked up on the strengths of domain-specific or task-specific models. Further learning how certain prompting methods may not work, or when done right would in fact improve prediction results was also eyeopening, as while I have learned about different prompting methods, looking at actual evaluation data from research gave a more concrete sense of how large this difference might be.

Papers 3 and 4 further broadened my knowledge and understanding of emotion recognition and mental health detection, introducing concepts of (utterance) emotion dynamics and properly integrating a psychological concept such as dynamic personality into a novel NLP framework that improved on two out of three datasets against existing state-of-the-art baselines. This better allowed me to now imagine what else we could possibly integrate into our models and frameworks, and how abstract concepts in psychology can be made concrete and functional with the right frameworks in NLP and datasets.

For example, imagine now if we had the resources to fine-tune a state-of-the-art LLM on domain-specific utterances data related to emotion recognition, personality classification, and mental health detection, further experimented with prompting methods perhaps beyond the most effective few-shot emotion-enhanced prompting we have seen in this review, and of course done so with

temporal ordering of data, and so on. We could even integrate other psychological concepts such as social dynamics within workspace setting, or physiological state of an individual eg. if they may be hungry (for example glucose levels, hormone), or feeling too hot or too cold, other more abstract concepts such as an individual's worldview and values, or even just environmental contexts such as the weather or how busy the streets the individual just passed by were. As abstract or implausible as it may sound, if such related data were made available, it would appear we just need to make use of them and improve our methods to achieve ever more effective models in emotion recognition, mental state/health recognition/detection, and the like, in utterances in texts or also other modalities.

References

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Will E. Hipson and Saif M. Mohammad. 2021. Emotion dynamics in movie dialogues. In *PLOS ONE*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE international conference on computer vision*.

James W. Pennebaker and Laura A. King. 1999. Linguistic styles: language use as an individual difference. In *Journal of personality and social psychology*.

Daniela Teodorescu, Tiffany Cheng, Alona Fyshe, and Mohammad Saif. 2023. Language and mental health: Measures of emotion dynamics from text as linguistic biosocial markers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Krishnapriya Vishnubhotla and Saif M. Mohammad. 2022. Tweet emotion dynamics: Emotion word usage in tweets from us and canada. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. Emotion recognition in conversation via dynamic personality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art

natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.