Multi-Task Learning for Depression Detection in Dialogs

Chuyuan Li[1], Chloé Braud[2], Maxime Amblard[1]
[1] Universite de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2] IRIT, Université de Toulouse, CNRS, ANITI, Toulouse, France
[1] {firstname.name}@loria.fr, [2] chloe.braud@irit.fr

Abstract

Depression is a serious mental illness that impacts the way people communicate, especially through their emotions, and, allegedly, the way they interact with others. This work examines depression signals in dialogs, a less studied setting that suffers from data sparsity. We hypothesize that depression and emotion can inform each other, and we propose to explore the influence of dialog structure through topic and dialog act prediction. We investigate a Multi-Task Learning (MTL) approach, where all tasks mentioned above are learned jointly with dialog-tailored hierarchical modeling. We experiment on the DAIC and DailyDialog corpora – both contain dialogs in English – and show important improvements over state-of-the-art on depression detection (at best 70.6% F_1), which demonstrates the correlation of depression with emotion and dialog organization and the power of MTL to leverage information from different sources.

1 Introduction

Depression is a serious mental disorder that affects around 5% of adults worldwide.[1] It comes with multiple causes and symptoms, leading to major disability, but is often hard to diagnose, with about half the cases not detected by primary care physicians (Cepoiu et al., 2008). Automated detection of depression, sometimes associated to other mental health disorders, has been the topic of several studies recently, with a particular focus on social media data and online forums (Coppersmith et al., 2015; Benton et al., 2017; Guntuku et al., 2017; Yates et al., 2017; Song et al., 2018; Akhtar et al., 2019; Ríssola et al., 2021). The ultimate goal of such system would be to complement expert assessments, but such empirical studies are also valuable to better understand how communication is affected by health disorders. In this paper, we propose to investigate depression detection within dialogs, a scenario less studied but more similar to the interviews with clinicians, which allegedly involves dialog features and also allows to examine how interaction is affected.

However, depression detection suffers from data sparsity. In fact, using social media data was a way to tackle this issue, including considering data generated by self-diagnosed users – a method that leads to potentially noisy data and comes with ethical issues (Chancellor et al., 2019). We rather examine a dataset of 189 clinical interviews, the DAIC-WOZ (Gratch et al., 2014), collected by experts to support the diagnosis of distress conditions. Participants are identified as depressive or not, and if so they receive a severity score. A line of work proposed to overcome data scarcity by leveraging varied modalities, e.g., using audio as in Al Hanai et al. (2018). Previous approaches were solely based on textual information relied on hierarchical contextual attention networks on word and sentence-level representations (Mallol-Ragolta et al., 2019), or Multi-Task Learning (MTL) but limited to combing identification and severity prediction (Qureshi et al., 2019; Dinkel et al., 2019), possibly with emotion (Qureshi et al., 2020).

Inspired by the latter approaches, we also propose relying on the MTL framework to help our model leverage information from different sources. We exploit three auxiliary tasks: emotion classification – naturally tied to mental health states –, and dialog act and topic classification, hoping the shallow information about the dialog structure could further enhance the performance. Our architecture is classic, based on hard-parameter sharing (Ruder, 2017), simpler than the shared-private architecture in (Qureshi et al., 2020) but has shown effective. In order to take into account dialog organization, we advocate for a dialog-tailored hierarchical architecture with some tasks performed at the speech turn level and others at the document level.

[1] https://www.who.int/news-room/fact-sheets/detail/depression

Proceedings of the SIGdial 2022 Conference, pages 68–75
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022 ©2022 Association for Computational Linguistics

Our contributions are: (i) An empirical study on depression detection in dialogs, leveraging the power of multi-task learning to deal with data sparsity; (ii) An extension of previous work in examining the effects of depression on dialog structure via shallow markers, i.e., dialog acts and topics, as a first step; (iii) State-of-the-art results on depression detection in DAIC test set with 70.6% in F_1 at best.

2 Related work

Within multi-task learning (MTL), a model has to learn shared representations to generalize the target task better. It improves the performance over single-task learning (STL) by leveraging commonalities or correlations between tasks. Recent years have witnessed a series of successful applications in various NLP tasks, as in Collobert and Weston (2008); Søgaard and Goldberg (2016); Ruder (2017); Ruder et al. (2019), which demonstrate the effectiveness of MTL in learning information from different but related sources. It also tackles the data sparsity issue and reduces the risk of overfitting (Mishra et al., 2017; Benton et al., 2017; Bingel and Søgaard, 2017).

Joshi et al. (2019) demonstrated the benefit of MTL for specific pairs of close health prediction tasks on tweets. Benton et al. (2017) used MTL on social media data and achieved important improvements in predicting several mental health signals, including suicide risks, depression, and anxiety, together with gender prediction. With a focus on depression detection, the shared task AVEC in 2016 (Valstar et al., 2016) has brought out a series of multi-modal studies using vocal and visual features on the DAIC-WOZ dataset (Gratch et al., 2014). Some of which also explored text-level features: Williamson et al. (2016) used Gaussian Staircase Model with semantic content features and reported a SOTA score on the validation set. Al Hanai et al. (2018) and Haque et al. (2018) learned sentence embeddings with an LSTM network. However, their results on textual features are lower than SOTA by a large margin. Dinkel et al. (2019) compared different word and sentence embeddings and various pooling strategies. Their best model is mean pooling with ELMo embeddings. Qureshi et al. (2019, 2020) proposed MTL approaches in adding emotion intensity and depression severity (i.e., a regression problem) prediction to the main classification task. They, however, found that the emotion-unaware model obtained the best result. They used a monologue corpus for the emotion task, a domain bias that possibly harms the performance. On the contrary, we hypothesize that emotional information would benefit depression detection. Mallol-Ragolta et al. (2019) used a hierarchical contextual attention network with static word embeddings within a single-task setting and then combined representations at the word and sentence levels. They reported at best 63% in F_1. Recently, Xezonaki et al. (2020) presented even better results, 70% in F_1, by augmenting the attention network with a conditioning mechanism based on effective external lexicons and incorporating the summary associated with each interview. We instead rely on MTL in this work, where incorporating external sources is more direct.

None of the previous studies investigated potential links between depression and dialog structure. We note that Cerisara et al. (2018) explored MTL with sentiment[2] and dialog act prediction on Mastodon (a Twitter-like dataset), where both annotations are available, and found a positive correlation. To the best of our knowledge, we are the first to tackle depression detection in dialog transcriptions with the MTL approach and explore joint learning techniques with tasks related to the dialog structure.

3 Model Architecture

One condition generally assumed for success within MTL, at least in NLP, is that the primary and auxiliary tasks should be related (Ruder, 2017). The emotion-related task is thus a natural choice since it is linked to mental states. We hypothesize that depressive disorder can also affect how people interact with others during conversations. We thus take a first step toward linking dialog structure and depression by examining shallow signals: dialog acts and topics. In addition, since the information comes at different levels, we propose hierarchical modeling, from speech turns to documents.

Baseline Model: Our basic model is a two-level recurrent network, similar to the one in Cerisara et al. (2018). The input words are mapped to vectors using word embeddings from scratch. The first level (*turn*-level) takes the embeddings into a bi-

[2] Sentiment and emotion are closely related with different function and/or granularity, cf. Munezero et al. (2014). Cerisara et al. (2018) use three labels for sentiment: *positive*, *negative*, *neutral*. In this paper, we use seven emotional labels: *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*, *neutral*.

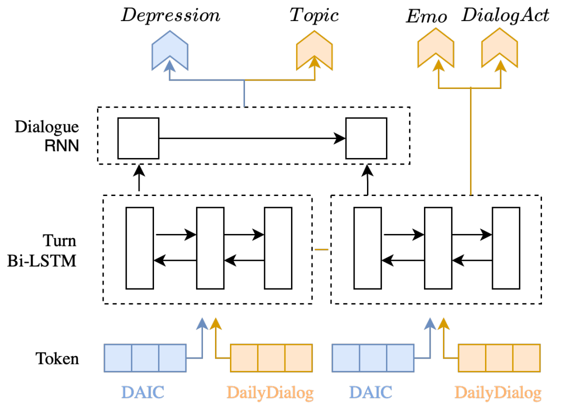

Figure 1: Multi-task fully shared hierarchical structure. Light blue is for DAIC dataset and depression task; orange is for DailyDialog and three auxiliary tasks.

LSTM network to obtain one vector for each turn. The second level (*dialog*-level) takes a sequence of turns into an RNN network, and the output is finally passed into a linear layer for depression prediction.

MTL Model: The MTL architecture is composed of shared hidden layers and task-specific output layers (see Fig. 1) and corresponds to the hard parameter sharing approach (Caruana, 1993, 1997; Ruder, 2017). Since some auxiliary tasks are annotated at the speech-turn level (i.e., emotion, dialog act) while others document level (i.e., depression, topic), our architecture is hierarchical and arranges task-specific output layers (MLP) at two levels. Sentence level emotion and dialog act information can be learned in the *turn*-level LSTM network and transferred upwards to help depression and topic prediction. On the other hand, higher-level information can be backpropagated to update the network at the lower level. The loss is simply the sum of the losses for each task. Regarding the MTL setting, we set equal weight for each task as the standard choice.

4 Datasets

DAIC-WOZ: This dataset is a subset of the DAIC corpus (Gratch et al., 2014).[3] It contains 189 sessions (one session is one dialog with avg. 250 speech turns) of two-party interviews between participants and Ellie – an animated virtual interviewer controlled by two humans. Table 1 gives the partition of train (107), development (35), and test (47) sets. Originally, patients are associated with a score related to the Patient Health Questionnaire (PHQ-9): a patient is considered depressive if PHQ-9 ≥ 10 (Kroenke and Spitzer, 2002).

	Train	Dev	Test
Depressed	77	23	33
Non Depressed	30	12	14
Total	107	35	47

Table 1: Number of sessions (dialogs) in DAIC-WOZ.

DailyDialog: This dataset (Li et al., 2017) contains $13,118$ two-party dialogs (with averaged 7.9 speech turns per dialog) for English learners,[4] covering various topics from ordinary life to finance. Three expert-annotated information are provided: 7 emotions (Ekman, 1999), 4 coarse-grain dialog acts, and 10 topics. We select this corpus due to its large size, two-level annotations and high quality. The train set contains $> 87k$ turns for emotions and dialog acts and $> 11k$ dialogs for topics. Detailed statistics are given in Appendix A.

5 Experimental setup

Baselines: We compare our MTL results with: (1) Majority class where the model predicts all positive; (2) Baseline single-task model (see Sec. 3); (3) State-of-the-art results on test set reported by Mallol-Ragolta et al. (2019) and Xezonaki et al. (2020). We do not compare to (Williamson et al., 2016; Haque et al., 2018; Al Hanai et al., 2018; Dinkel et al., 2019; Qureshi et al., 2020) who only report on the development set.

Evaluation Metrics: For depression classification we follow Dinkel et al. (2019) and report accuracy, macro-F$_1$, precision, and recall. For emotion analysis, we follow Cerisara et al. (2018) and report macro-F$_1$.

Implementation Details: We implement our model with AllenNLP library (Gardner et al., 2018). We use the original separation of train, validation, and test sets for both corpora.

The model is trained for a maximum of 100 epochs with early stopping. For STL as well as for MTL scenario, we optimize on macro-F$_1$ metric for depression classification. We use cross-entropy loss. The batch size is 4 for DailyDialog and 1 for DAIC (within the limit of GPU VRAM). We

[3] https://dcapswoz.ict.usc.edu

[4] http://yanran.li/dailydialog

use the tokenizer from spaCy Library (Honnibal et al., 2020) and construct the word embeddings by default with a dimension of 128. The *turn* level has one hidden layer and 128 output neurons. We tune *document* RNN layers in $\{1, 2, 3\}$ and hidden size in $\{128, 256, 512\}$. Model parameters are optimized using Adam (Kingma and Ba, 2014) with $1e-3$ learning rate. Dropout rate is set to 0.1 for both *turn* and *document* encoders. The source code is available at `https://github.com/chuyuanli/MTL4Depr`.

6 Results and Discussion

6.1 Depression Detection Results on DAIC

Results using MTL hierarchical structure are shown in Table 2, which are compared to majority vote and SOTA models (at the top). Our baseline model is a single-task naive hierarchical model which obtains similar results (F_1 44) as the baseline model (NHN) in Mallol-Ragolta et al. (2019) (F_1 45).

Using the multi-task architecture, we get improvements when adding each task separately. We see more than a +11.5% increase in F_1 when adding emotion ('+Emo') or topic ('+Top') classification task and, at best, +16.9% with dialog acts ('+Diag'). This demonstrates the relevance of each task to the primary problem of depression detection, especially the interest of dialog acts. When adding topics, we observe a small drop in accuracy compared to STL while the F_1 is better, meaning that the prediction for minority class (non-depressive) improves. Interestingly, in terms of accuracy, the tasks at different levels (depression '+Emo' and depression '+Diag') seem to help more. We deduce that they help build a better local representation (speech turns) before the global representation.

When jointly learning all four tasks – combining depression detection with three auxiliary tasks ('+Emo+Diag+Top') –, all metrics improve. We obtain our best system with an improvement of +26.7% in F_1 compared to STL baseline, outperforming the state-of-the-art with a +7.6% increase compared to the best system in Mallol-Ragolta et al. (2019) and about +0.5% compared to Xezonaki et al. (2020). Depressed people tend to express specific emotions; it is thus natural to think that emotion is beneficial for the main task. These results indicate that both emotion and dialog structure help as they provide complementary information, paving the way for new research directions with more fine-grained modeling of dialog structure for tasks in conversational scenarios.

	F_1	Prec.	Rec.	Acc.
BSL Majority vote	41.3	35.1	50.0	70.2
State-of-the-art				
NHN[5] (Mallol-Ragolta et al., 2019)	45	-	50	-
HCAN[6] (Mallol-Ragolta et al., 2019)	63	-	66	-
HAN+L[7] (Xezonaki et al., 2020)	70	-	70	-
Ours				
STL Depression	43.9	44.5	47.5	63.8
MTL +Emo	55.5	56.2	61.6	70.2
MTL +Top	55.6	55.9	56.8	59.6
MTL +Diag	60.8	60.6	61.4	66.0
MTL +Emo+Diag+Top	**70.6***	**70.1**	**71.5***	**74.5**

Table 2: Depression detection results on DAIC. STL: single-task using DAIC only; MTL: multi-task using DAIC and adding classification for Emotion (+Emo), Topic (+Top), Dialog Act (+Diag) from DailyDialog. *Significantly better than SOTA performance with p-value < 0.05.

6.2 Analysis

Performance on Auxiliary Tasks: To better understand our model, we look at the performance of emotion, dialog act, and topic auxiliary tasks. Directly comparing the results of our MTL approach ('+Emo+Diag+Top') with a STL architecture for each task, however, seems unfair. The optimized objective and structural complexity are different: the former is optimized on the depression detection task on two levels, while the latter is tuned on the target auxiliary task with either speech turn (emotion and dialog act) or full dialog (topic). Unsurprisingly, the results show that the MTL system underperforms the basic STL structure for dialog acts and topics, with at best 67.8 in F1 (MTL) *vs.* 68.8 (STL) for dialog acts, and 52.0 (MTL) *vs.* 52.4 (STL) for topic classification.

For emotion, on the other hand, our best MTL system obtains 40.0 in F_1 compared to 38.3 for the STL baseline, showing the mutual benefit of both tasks. Even though the score is lower than the SOTA for emotion classification (51.0 F_1 in Qin et al. (2021))[8], we believe that refining our model for this task could lead to further improvements in depression detection. In addition, we observe that our MTL approach is particularly beneficial for negative and rare emotion classes, with *anger*,

[5] Naive hierarchical network (baseline).

[6] Hierarchical contextual attention network.

[7] Hierarchical attention network with LIWC lexicon.

[8] Precision: in Qin et al. (2021) authors report results on sentiment classification. It is yet unclear how they convert emotion annotation (7 labels) to sentiment (3 labels).

High-level DA	#	%	Sub-cat.	#	%
Question	$7,907$	53%	Emo	$1,054$	13%
			Non-emo	$6,853$	87%
Backchannel	$3,231$	22%	-	-	-
Comment	$3,074$	20%	-	-	-
Opening	611	4%	-	-	-
Other	171	1%	-	-	-

Table 3: High-level dialog act distribution of Ellie in DAIC-WOZ. # and % represent the number and percentage of Ellie's utterances, respectively.

disgust and *sadness* gaining resp. 5%, 6% and 1% in F_1. Finally, we conduct a manual inspection of the types of utterances (mostly questions) from Ellie, and classify them into high-level dialog acts: *Backchannel, Comment, Opening, Other, Question*.[9] We find that around 13% of the utterances are emotion-related, for instance "things which make you mad / you feel guilty about, last time feel really happy", etc., and that mentions of topics related to happiness or regret appear in almost all the interviews. Dialog act distribution is shown in Table 3. We release our annotation to the community for future studies.

Effectiveness of Hierarchical Structure: To examine the effectiveness of hierarchical structure, we conduct ablation studies on the full multi-learning setting ('+Emo+Diag+Top'). For dialog RNN level, we use topic information; for turn level, we test either emotion or dialog act. The results are shown in Table 4. Unsurprisingly, both ablated models ('+Emo+Top' and '+Diag+Top') underperform the full model, with F_1 scores decreasing $\approx 6\%$ each. Without dialog act, all metrics drop, showing the importance of this information for dialog structure. Without emotion, recall drops dramatically while accuracy and precision increase, indicating that the model '+Diag+Top' predicts more positive classes but fails in negative ones, which could result in too many false positives in real-life scenarios. On the other hand, when comparing hierarchical models ('+Emo+Top', '+Diag+Top', '+Emo+Diag+Top') with single-level models ('+Emo', '+Top', '+Diag'), we see considerable improvements in all metrics, and this holds for all auxiliary tasks. We can thus confirm the advantage of hierarchical structure for model performance.

		F_1	Prec.	Rec.	Acc.
MTL	+Emo+Diag+Top	**70.6**	70.1	**71.5**	74.5
MTL	+Emo+Top	64.4	64.4	64.4	70.2
MTL	+Diag+Top	63.7	**78.1**	62.8	**76.6**

Table 4: Ablation study on hierarchical structure.

7 Conclusion

In this paper, we demonstrate the correlation between depression and emotion and show the relevance of features linked to dialog structures via shallow markers: dialog acts and topics. In the near future, we intend to investigate more refined modeling of dialog structures, possibly relying on discourse parsing (Shi and Huang, 2019). We would also like to explore depression severity classification as an extension to binary classification, possibly through a cascading structure: first, detect depression and then classify the severity. We intend to refine our work and report on cross-validation splits of the data to test the stability of the model, an issue even more crucial when dealing with sparse data with possibly representativeness problem. A further step will be to investigate the generalization of our model to other mental health disorders.

Acknowledgement

The authors thank the anonymous reviewers for their insightful comments and suggestions. This work was supported by the PIA project "Lorraine Université d'Excellence", ANR-15-IDEX-04-LUE, as well as the CPER LCHN (Contrat de Plan État-Région - Langues, Connaissances et Humanités Numériques). It was also partially supported by the ANR (ANR-19-PI3A-0004) through the AI Interdisciplinary Institute, ANITI, as a part of France's "Investing for the Future — PIA3" program, and through the project AnDiAMO (ANR-21-CE23-0020). Experiments presented were carried out in secured clusters on the Grid'5000 testbed. We would like to thank the Grid'5000 community (`https://www.grid5000.fr`).

Ethical Considerations

The goal of such systems is not to replace human healthcare providers. All these systems may be used only in support to human decision. The principle of leaving the decision to the machine would imply major risks for decision making in the health field, a mistake that in high-stakes healthcare set-

[9] *Backchannel* refers to phatic expressions such as *yeah, hum mm*. Here we use different dialog acts from those in DailyDialog.

tings could prove detrimental or even dangerous.

Another issue is the representativeness of the data. Currently, it is very complex to access patients in order to have more examples. The institutional complexity leads researchers to systematically use the same data set, creating a bias between the representation of the pathology, in particular for mental ones whose expression can take very varied forms. This also implies defining a variation in relation to a normative use of language that comes with a strong risk in this type of approach.

Moreover, we carefully select the dialog corpora used in this paper to control for potential biases and personal information leakage. We only work with interview transcription, with no audio or visual information. For the text part, all the participant's name have been marked out with pseudo-ID.

References

Shad Akhtar, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya, and Sadao Kurohashi. 2019. All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE transactions on affective computing*.

Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. 2018. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Rich Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Machine learning: Proceedings of the tenth international conference*, pages 41–48.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Monica Cepoiu, Jane McCusker, Martin G Cole, Maida Sewitch, Eric Belzile, and Antonio Ciampi. 2008. Recognition of depression by non-psychiatric physicians—a systematic literature review and meta-analysis. *Journal of general internal medicine*, 23(1):25–36.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 745–754, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 79–88.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39.

Heinrich Dinkel, Mengyue Wu, and Kai Yu. 2019. Text-based depression detection on sparse data. *arXiv preprint arXiv:1904.05154*.

Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Albert Haque, Michelle Guo, Adam S Miner, and Li Fei-Fei. 2018. Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. 2019. Does multi-task learning always help?: An evaluation on health informatics. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 151–158, Sydney, Australia. Australasian Language Technology Association.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kurt Kroenke and Robert L Spitzer. 2002. The phq-9: a new depression diagnostic and severity measure.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn W Schuller. 2019. A hierarchical attention network-based approach for depression detection from transcribed clinical interviews. *Proc. Interspeech 2019*, pages 221–225.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–387, Vancouver, Canada. Association for Computational Linguistics.

Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2021. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In *AAAI*.

Syed Arbaaz Qureshi, Gaël Dias, Mohammed Hasanuzzaman, and Sriparna Saha. 2020. Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine*, 15(3):47–59.

Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. 2019. Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, 34(5):45–52.

Esteban A Ríssola, David E Losada, and Fabio Crestani. 2021. A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare*, 2(2):1–31.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv e-prints*, pages arXiv–1706.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. Latent multi-task architecture learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4822–4829.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7007–7014.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany. Association for Computational Linguistics.

Hoyun Song, Jinseon You, Jin-Woo Chung, and Jong C. Park. 2018. Feature attention network: Interpretable depression detection from social media. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10.

James R Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F Quatieri. 2016. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 11–18.

Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth Narayanan. 2020. Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. In *INTERSPEECH*, pages 4556–4560.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

A Auxiliary Tasks Class Distribution in DailyDialog

Table 5, Table 6, and Table 7 show the number and percentage of emotion, dialog act, topic for each subset, resp.

Emotion	Train		Dev		Test	
	#	%	#	%	#	%
0-no emotion	72, 143	82.8	7, 108	88.1	6, 321	81.7
1-anger	827	0.9	77	1.0	118	1.5
2-disgust	303	0.3	3	0.04	47	0.6
3-fear	146	0.2	11	0.1	17	0.2
4-happiness	11, 182	12.8	684	8.5	1019	13.2
5-sadness	969	1.1	79	1.0	102	1.3
6-surprise	1, 600	1.8	107	1.3	116	1.5
Utt. Total	87, 170	100.0	8, 069	100.0	7, 740	100.0

Table 5: Emotion distribution in train, dev. and test sets.

Dialog Act	Train		Dev		Test	
	#	%	#	%	#	%
1-inform	39, 873	45.7	3, 125	38.7	3, 534	45.7
2-question	24, 974	28.6	2, 244	27.8	2, 210	28.6
3-directive	12, 242	16.3	1, 775	22.0	1, 278	16.5
4-commissive	8, 081	9.23	925	11.5	718	9.3
Utt. Total	87, 170	100.0	8, 069	100.0	7, 740	100.0

Table 6: Dialog act distribution in train, dev. and test sets.

Topic	Train		Dev		Test	
	#	%	#	%	#	%
1-ordinary life	2, 975	26.8	418	41.8	252	25.2
2-school life	453	4.1	0	0	34	3.4
3-culture & education	50	0	0	0.0	5	0.5
4-attitude & emotion	616	5.5	1	0.0	50	0.5
5-relationship	3, 879	34.9	129	12.9	384	38.4
6-tourism	860	7.7	124	12.4	79	7.9
7-health	205	1.8	41	4.1	21	2.1
8-work	1, 574	14.2	215	21.5	135	1.4
9-politics	105	0.9	13	1.3	13	1.3
10-finance	399	3.6	59	5.9	27	2.7
Total	11, 118	100.0	1, 000	100.0	1, 000	100.0

Table 7: Topic distribution in train, dev. and test sets.