Enhancing Depression Detection from Narrative Interviews using Language Models

Palak Sood, Xinming Yang, Ping Wang
Department of Computer Science
Stevens Institute of Technology
Hoboken, New Jersey, USA

Outline

- **Introduction**
- **Existing Challenges**
- **Our Contributions**
- **Experimental Results**
- **Conclusion**

Image source

Introduction

- Mental health in our society is **declining**. Depression is a common and serious mental health issue in our society.
- According to a 2020 report by the National Institute of Mental Health (NIMH),
  - Nearly 53 million U.S. Americans (21% of all adults) suffer from, or experience, some form of mental illness.
  - The reported prevalence of any mental illness was highest for adults reporting two or more races/ethnicities.
  - Higher for female respondents than for male respondents.
  - A growing percentage of adolescents in the US live with major depression.
  - The prevalence of mental illness is highest (31%) for those younger than 25 years (67% of college students).

Pester, Christian W., Gina Noh, and Andi Fu. "On the Importance of Mental Health in STEM." *ACS Polymers Au* (June 2023).

Urgent Needs for Automatic Early Detection

- **Final Objectives**: Prompt intervention; improve student outcomes, reduce risks of further deterioration
- **Current status**:
  - Many online mental health tools to provide personalized support and consultations.
  - Some datasets are released to facilitate the task
    - Surveys, interviews, online platforms
    - Providing valuable insights into the underlying patterns and characteristics of different mental health conditions.

Foundational resources for developing effective machine learning (ML) and natural language processing (NLP) models for mental health issue detection.

Existing Challenges

- **Data scarcity:**
  - Privacy concerns hinder the collection of large datasets
- **Label imbalance:**
  - Limited data collection and imbalanced depression instances
  - Limiting the development of accurate models for depression detection
- **Long text:**
  - Analyzing long text inputs in interviews is challenging
- **Sufficient contexts:**
  - Traditional machine learning methods for depression detection often rely on word frequency
  - Overlooking informative contextual dependencies

Our Contributions

Enhancing depression detection from narrative interviews using language models:

- Create an integrated interview corpus named I-DAIC by combining three existing datasets.
- Conduct a comprehensive evaluation of two pre-trained language models on depression detection by comparing with two traditional machine learning methods.
- Investigate several customized strategies for handling long text inputs about the narrative interviews.
- Identify representative keywords for both depression and non-depression instances with topic modeling.

Data Integration

- To overcome the data scarcity issue
- Collected three datasets:
  - DAIC WOZ (English)
  - Extended DAIC WOZ (English)
  - EATD corpus (Chinese)
- Utilize machine translation to translate all into English.
- Integrate them to one dataset, named Integrated I-DAIC Dataset.
- The more the data, the better a model can potentially understand and analyze aspects of mental health.

Data	DAIC	E-DAIC	EATD	I-DAIC
Train Total	107	163	129	399
Non-depression	77	126	105	308
Depression	30	37	24	91
Dev Total	35	56	17	108
Non-depression	23	44	14	81
Depression	12	12	3	27
Test Total	47	56	16	119
Non-depression	33	39	13	85
Depression	14	17	3	34

	Training	Development	Testing
avg_words	1,116.04	1,257.94	1,360.70
min_words	9	0	0
max_words	4,622	3,440	5,011
avg_sentences	70.44	84.11	92.51
min_sentences	2	1	1
max_sentences	209	204	197

Statistics of the Integrated I-DAIC dataset.

An Interview Example

Conversation -
Ellie - how would your best friend describe you .
Patient - i'm a good friend i'm a true friend i'm honest i'm real i'm dependable and i don't play games i'm very i'm no <n> i'm no drama .
Ellie - have you ever served in the military . **Patient** - no.
Ellie - have you ever been diagnosed with p_t_s_d . **Patient** - yes i have.
Ellie - how long ago were you diagnosed .
Patient - in um february of two thousand eleven .
Ellie - what got you to seek help .
Patient - i was attacked by a stalker and almost killed in november of two thousand nine he broke into my apartment and laid in wait for me and attacked me when i came in the door and tried to kill me .
Ellie - do you still go to therapy now . **Patient** - i do .
Ellie - how easy is it for you to get a good night's sleep .
Patient - it's not it's never easy it's always bad .
Ellie - what are you like when you don't sleep well .
Patient - tired lethargic um it's hard to keep my thoughts in order it's hard just to do the basics during the day .
Ellie - are they triggered by something . **Patient** - mm no.
Ellie - is there anything you regret.
Patient- i have too many regrets right now .

Label - 1 (depressed)

Automatic Depression Detection Models

Traditional Machine Learning (ML) Methods

- Require significant feature engineering based on the frequency of words
- Ignore the semantic relationships between words and their contextual information.
- Models used:
  - **SVM** (Support Vector Machine)
  - **LR** (Logistic Regression)

Task-Specific Fine-tuned Language Models

- We fine-tuned two Transformer-based pre-trained language models on the specific depression detection task on the I-DAIC dataset.
  - **BERT** (Bidirectional Encoder Representations from Transformers)
  - **RoBERTa** (Robustly Optimized BERT)
- Can capture rich contextual information and semantic relationships.

Strategies to Handle Long Text

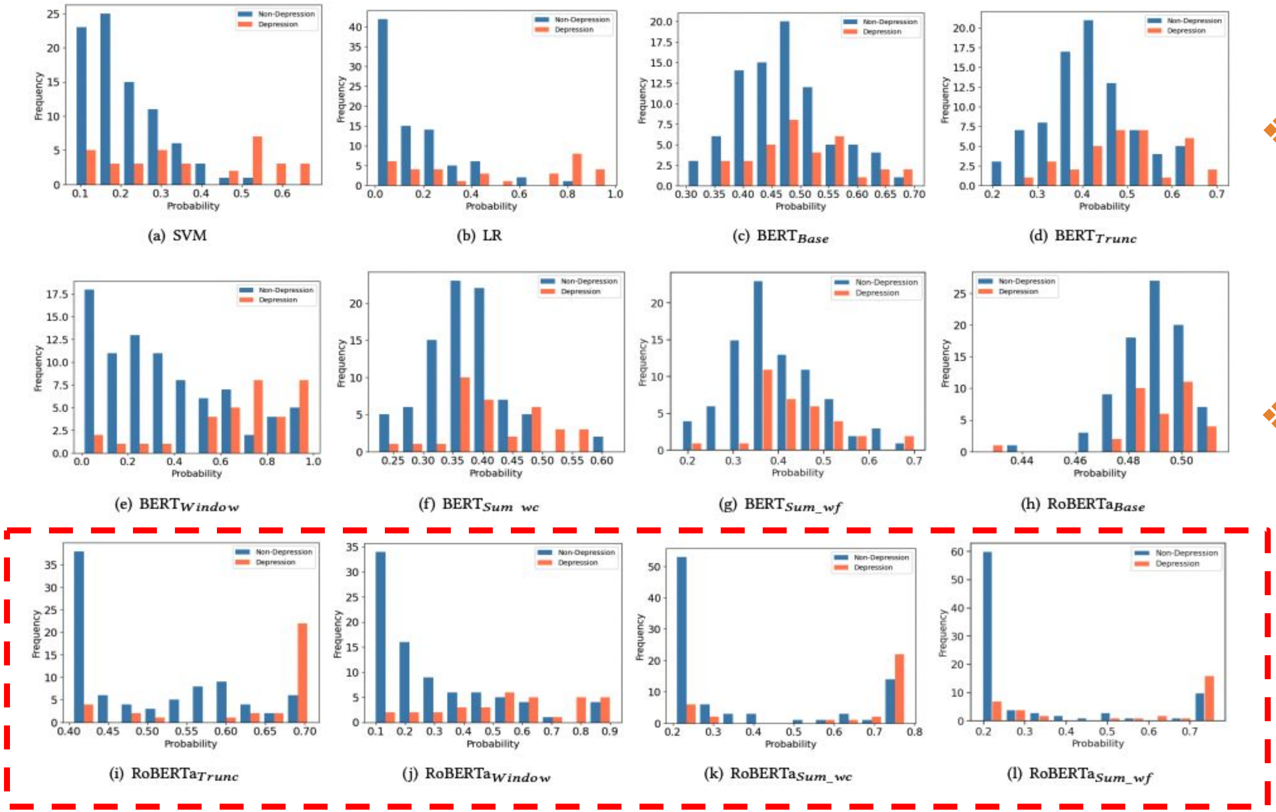
- **Customized Truncation (*Trunc*)**:
  - Capture the middle part of a conversation, ignore the introductory part
  - Ensure important dialogues are considered for depression detection
- **Sliding Window (*Window*)**:
  - Divide the sample text into overlapping 512-token chunks, and process them individually
  - Average the predicted probabilities to obtain the combined probability for the entire text
- **Extractive Summarization by Word Count (*Sum_wc*)**:
  - Sentences with higher word counts are selected for the final summary
  - Retain longer sentences with richer information
- **Extractive Summarization by Word Frequency (*Sum_wf*)**:
  - Sentence rating is based on the frequencies of the words in each sentence
  - Retain a compressed text that captures the conversation's general topics

Overall Performance on Depression Detection

Model	Non-Depression			Depression			Weighted			KL Divergence
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	
SVM	0.84	0.94	0.89	0.79	0.56	0.66	0.83	0.83	0.82	0.0357
LR	0.82	0.96	0.89	0.84	0.47	0.60	0.83	0.82	0.81	0.0324
BERT$_{Base}$	0.75	0.68	0.72	0.36	0.44	0.39	0.64	0.61	0.62	0.0526
BERT$_{Trunc}$	0.79	0.81	0.80	0.50	0.47	0.48	0.71	0.71	0.71	0.0461
BERT$_{Window}$	0.78	0.88	0.83	0.57	0.38	0.46	0.72	0.74	0.72	0.0417
BERT$_{Sum_wc}$	0.75	**0.98**	**0.85**	**0.75**	0.18	0.29	0.75	0.75	0.69	0.0441
BERT$_{Sum_wf}$	0.73	0.85	0.79	0.38	0.24	0.29	0.63	0.67	0.65	0.0466
RoBERTa$_{Base}$	0.73	0.82	0.77	0.35	0.24	0.28	0.62	0.66	0.63	0.0545
RoBERTa$_{Trunc}$	**0.89**	0.56	0.69	0.43	**0.82**	0.57	0.76	0.64	0.65	0.0503
RoBERTa$_{Window}$	0.75	0.94	0.83	0.58	0.21	0.30	0.70	0.73	0.68	**0.0352**
RoBERTa$_{Sum_wc}$	**0.89**	0.76	0.82	0.57	0.76	**0.65**	**0.80**	0.76	0.77	0.0390
RoBERTa$_{Sum_wf}$	0.85	0.85	**0.85**	0.62	0.62	0.62	0.78	**0.78**	**0.78**	0.0376

The results demonstrated the effectiveness, advantage, and potential of advanced language models for depression detection on interview corpora.

Evaluation of Discriminative Capability



- The distribution of predicted scores
  - Depression (in orange)
  - Non-depression (in blue)
- We can observe large separation between the two classes when using RoBERTa Summ Models.

Representative Keywords with Topic Modeling

Keywords for Depression

Keywords for Non-Depression

Class	Representative Keywords
Depression	Dispute, Depressed, Work, Semester, Unemployed, Sleep, Jobs, Goodbye
Non-Depression	Cute, Happy, Friends, Attracted, Unhappy, Future, Learning, Exploring

Conclusion

- This study aims to enhance depression detection from narrative interviews using language models.
- **Our contributions:**
  - Data integration to get a comprehensive dataset
  - Fine-tuning and evaluation of two Transformer-based pre-trained language models
  - Investigation of several strategies for handling long text
  - Identification of representative keywords for depression and non-depression
- **Our findings:**
  - RoBERTa outperforms BERT in terms of efficiency
  - Summarization-based strategy works best for long-text inputs
  - Keywords tell us a lot about the depression and non-depression classes

Thank you!

Link to I-DAIC dataset and codes:
https://github.com/LEAF-Lab-Stevens/IDAIC

Feel free to send questions and suggestions to:

ping.wang@stevens.edu

psood@stevens.edu