Leveraging Mental Health Forums for User-level Depression Detection on Social Media

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Sravani Boinepelli, Tathagata Raha, Harika Abburi, Pulkit Parikh, Niyati Chhaya, Vasudeva Varma

Information Retrieval and Extraction Lab, IIIT Hyderabad, India

Motivation

- Applying depression detection mechanisms to social media platforms is challenging due to the vastness and inherent imbalance of content. We design a hierarchical neural framework to optimize the performance of user-level depression classification systems. To simulate a platform agnostic framework, we simultaneously replicate the size and composition of social media to identify victims of depression.
- Mental health blogs are valuable sources of linguistic cues that could help identify depression from texts while retaining user privacy. We introduce a novel dataset derived from such blogs to incorporate domain-specific knowledge into our framework.

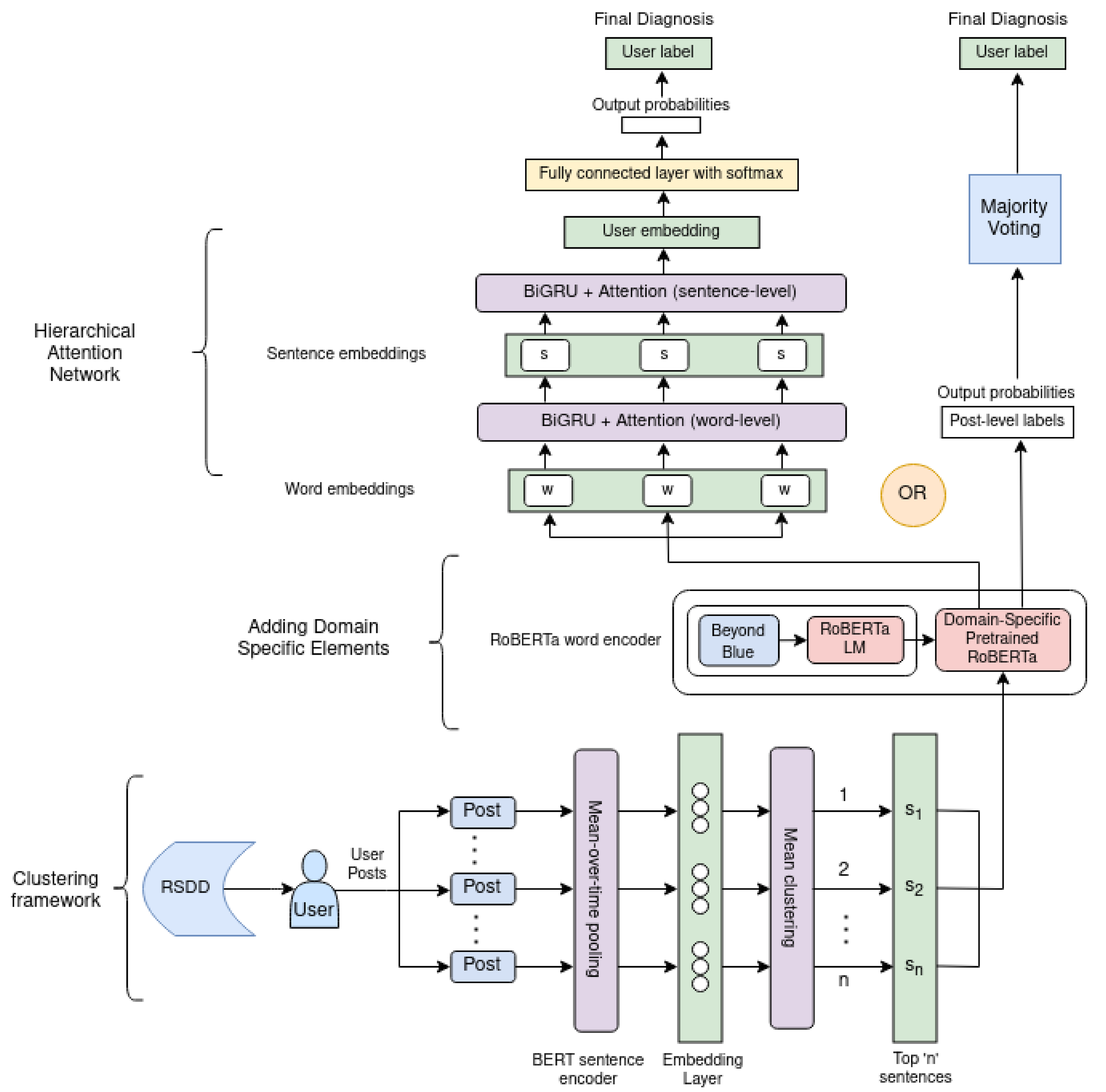
Dataset	Statistic	Count
Mental Health Blog Dataset	Mean length of posts	230
	# of Sentences	452756
	NER count	188259
	Total word count	4439028
	# of Unique words	49904
	Forms of Depression	7708
	Depression WordNet	359

The Mental Health Blog Dataset

- To collect data, mental health blogs were crawled for posts spanning a timeline from 31st December 2011 to 25th June 2020.
- This yielded a mental health language resource containing a total of 39248 posts with 9354 total unique users.
- The dataset also comes with other helpful information such as date of posting, number of posts previously posted by the author, number of likes received to mark how helpful a post was, etc.

Reddit Self-reported Depression Diagnosis (RSDD)

- We use the large-scale publicly available RSDD dataset which contains 9,210 depressed users and 107,274 non-depressed users, with an average of 969 posts for each user and a median of 646.



Proposed Methodology

- **Clustering framework:** Compute embeddings for all posts of each user using BERT as a sentence encoder. Then compute the mean of all the posts of a particular user. The top 'n' posts that are nearest to the mean are selected.
- **Domain-Specific Pre-training:** Entries from our MHB dataset are utilized to tailor a pre-trained RoBERTa model and obtain more effective representations for our model.
- **Tackling imbalance:** The ratio of positive to negative samples in the RSDD training data is 1:10.57. We therefore use a variation of cross-entropy loss that is weighted in nature.
- **Assigning User-level labels:**
  - Majority voting: Given that the posts we are considering represent the user's 'n' most relevant posts, we postulate that the 'degree' of depression in the user can be ascertained by simply taking the mode of the outputted 'n' post labels.
  - Hierarchical Attention Network: This allows us to pay variable attention to individual words and sentences in a context-specific manner when constructing representations for the user's posts.

Model	P	R	F1
BERT	0.54	0.51	0.52
XLNet	0.55	0.57	0.55
RoBERTa	0.55	0.56	0.56
X-A-BiLSTM	0.69	0.53	0.60
UserCNN-E	0.59	0.45	0.51
UserCNN-R	0.75	0.57	0.65
Domain-specific, MV	0.67	0.63	0.65
Domain-specific, HAN	**0.79**	**0.66**	**0.72**

Experimental Results

- Our clustering mechanism is memory and time efficient. It focuses on relevant features without arbitrarily losing vital user information. Our class weighted loss function improves class imbalance.
- Mechanisms such as majority voting have proven to disregard the nuances in misclassified posts while our hierarchical deep learning network effectively picks out important words or sentences from classified posts and ultimately allows us to label a user as depressed.
- The use of our mental health blog dataset contributes integral domain-specific information while simultaneously emulating the degree of informality with which such topics are discussed on social media. Tuning the pretrained representations of our Beyond Blue dataset to the larger Reddit dataset also outperform popular models and are known to improve model performance significantly.

Model	Acc	P	R	F1	Time/epoch
Post-level, Mean Clustering	**90.5**	**0.91**	0.33	0.48	71 min
Post-level, Clustered+CW(Class-weighted)	78.78	0.58	**0.70**	0.59	70 min
Post-level, Clustered+CW+Domain-specific	82.2	0.77	0.64	0.69	69min
Post-level, Unclustered+CW+Domain-specific	82.7	0.79	0.64	**0.70**	1343min
User-level, Majority Voting	83.3	0.67	0.63	0.65	69 min
User-level, HAN	**84.8**	**0.79**	**0.66**	**0.72**	180 min