NLP:

Depression Detection

Kanawut Kaewnoparat (R1)
Win Win Phyo (R2)
Lin Tun Naing (R3)

R1: Predicting Depression via Social Media

Objective: This paper aims at predicting the likelihood of depressive disorder by analyzing the **sociolinguistic patterns** of social media posts for a 1-year time before the reported onset of 476 participants (171 pos, 305 neg).

Finding: Decreased social activity, increased nocturnal engagement , raised negative affect, highly clustered egonetworks, depressing linguistic style, heightened relational and medical concerns and greater expression of religious involvement **>> signs of depressive symtoms**

Model: SVM with RBF kernel

Result: The accuracy of 70%, with precision at 74%

Further area: scalable method for automated public health tracking

values to be statistically significant.

	Mean	Variance	Momentum	Entropy
volume	15.21***	14.88***	14.65***	17.57***
replies	22.88***	13.89	29.18***	19.48***
questions	8.205	7.14	23.06***	10.71
PA	14.64	10.94	13.25	17.74***
NA	16.03***	19.01***	17.54***	15.44***
activation	19.4 ***	17.56***	22.49***	17.84***
dominance	20.2 ***	18.33***	24.49***	12.92
#followees	28.05***	14.65	25.95***	16.85***
reciprocity	5.24 ***	5.35	7.93 ***	6.82 ***
clust. coeff.	12.33***	10.92	15.28***	11.91
#ego comp.	7.29	6.91 ***	9.04 ***	8.56
antidepress	8.68	10.13	10.17***	5.73
depr. terms	22.29***	16.28***	22.16***	18.64***
1st pp.	25.07***	15.26***	24.22***	19.77***
2nd pp.	13.03***	12.43	20.36***	11.49
3rd pp.	20.34***	14.60	21.47***	16.96***
article	9.75	14.41	16.68***	7.60
negate	8.42	6.33	16.7 ***	12.13
swear	12.91	6.12	20.8 ***	18.99***

*** $p \leq \alpha$, after Bonferroni correction df=474

Table 5: Statistical significance (t-statistic values) of the mean, variance, momentum and entropy measures of selected dynamic features, comparing the depression and non-depression classes.

	precision	recall	acc. (+ve)	acc. (mean)
engagement	0.542	0.439	53.212%	55.328%
ego-network	0.627	0.495	58.375%	61.246%
emotion	0.642	0.523	61.249%	64.325%
linguist. style	0.683	0.576	65.124%	68.415%
dep. language	0.655	0.592	66.256%	69.244%
demographics	0.452	0.406	47.914%	51.323%
all features	0.705	0.614	68.247%	71.209%
dim. reduced	**0.742**	**0.629**	**70.351%**	**72.384%**

Table 6. Performance metrics in depression prediction in posts using various models. Third column shows the mean accuracy of predicting the positive class.

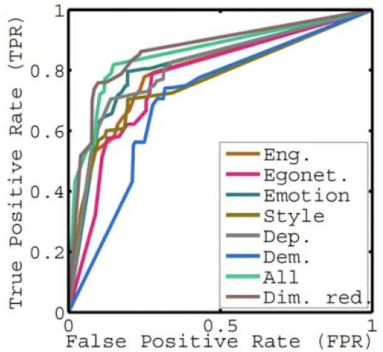


Figure 4. Receiver Operating Characteristic (ROC) curves in predicting labels of users. Each curve corresponds to a model trained on a particular feature type.

R2: Monitoring Depression Trend on Twitter during the COVID-19 Pandemic

Objective: This paper states that the model's capability of monitoring both tweet-chunk level and user level of depression scores during COVID-19 pandemic.

Data Collection: use Tweepy API to retrieve the depression related tweets with maximum 200 tweets per user within last 3 months by April 18 2020 , 2,575 distinct users for depression-related and
the another 2,575 for non-depression users.

- User level collections: Personality, Sentiments, Demographics, LIWC, Social Media Engagement
- Chunk level collections: concatenate consecutive tweets of the same user to create tweet chunks of 250 words and label the chunks

Models: Attention BiLSTM and multichannel CNN (baseline), transformer -based classification models(BERT, RoBERTa, XLNet), ML classification methods(random forest, logistics regression,SVM)

Experient: Accuracy of Chunk-level(XLNet - 77.1%)
Accuracy of User-level(SVM - 78.9%)

Model	Train-Val Set	Accuracy	F1	AUC	Precision	Recall
Attention BiLSTM	1k users	70.7	69.0	76.5	70.9	67.3
	2k users	70.3	68.3	77.4	70.7	66.1
	4.65k users	72.7	71.6	79.3	72.1	71.1
CNN	1k users	71.8	72.6	77.4	72.7	72.6
	2k users	72.8	74.5	80.3	72.2	76.9
	4.65k users	74.0	70.9	81.0	77.4	68.9
BERT	1k users	72.7	74.4	79.8	72.0	76.9
	2k users	75.7	76.3	82.9	76.1	75.7
	4.65k users	76.5	77.5	83.9	76.3	78.8
RoBERTa	1k users	74.4	75.7	82.0	74.2	77.3
	2k users	75.9	77.9	83.2	73.8	**82.5**
	4.65k users	76.2	**78.0**	84.1	74.4	81.9
XLNet	1k users	73.7	75.1	80.7	73.2	77.2
	2k users	74.6	76.8	82.6	72.6	81.5
	4.65k users	**77.1**	77.9	**84.4**	**77.5**	78.3

Table 1: Chunk-level performance (%) of all 5 different models using training-validation sets of different sizes.

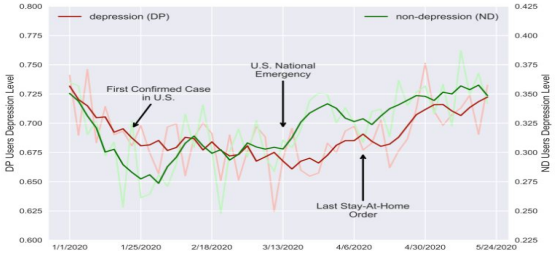
Features	Accuracy	F1	AUC
VADER	54.9	61.7	54.6
Demographics	58.7	56.0	61.4
Engagement	58.7	62.3	61.7
Personality	64.8	67.8	72.4
LIWC	70.6	70.8	76.0
V+D+E+P+L	71.5	72.0	78.3
XLNet	78.1	77.9	84.9
All (Rand. Forest)	78.4	78.1	84.9
All (Log. Reg.)	78.3	78.5	**86.4**
All (SVM)	**78.9**	**79.2**	86.1

Table 2: User-level performance (%) using different features. We use SVM for classifying individual features.

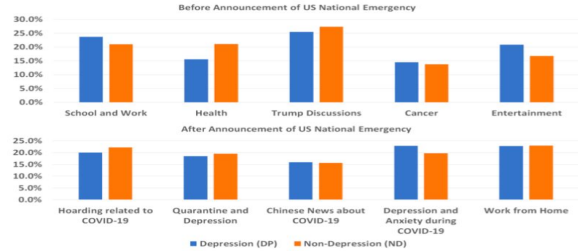
R2: Monitoring Depression Trend on Twitter during the COVID-19 Pandemic

Application: Two COVID-19 related applications of XLNet based depression classifier:

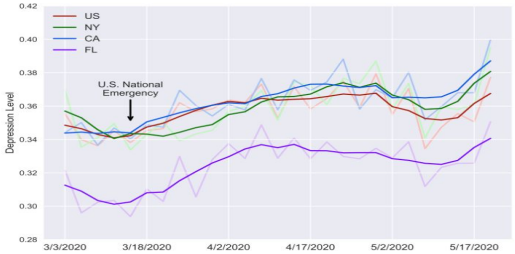
- Depression Monitoring on DP/ND Group
- Monitoring Depression Level at the U.S country level and state level during pandemic



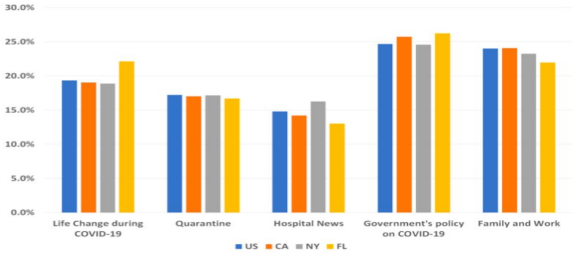
(a) DP-ND trends



(b) Percentage of DP-ND topics



(c) State-level trends



(d) Percentage of State-level topics

Figure 5: (a) Aggregated depression level trends of DP users and ND users from January 1st, 2020 to May 22nd, 2020. We use different y-axes for the 2 groups in order to compare them side by side. (b) Topics of DP and ND before and after the announcement of the U.S. National Emergency. (c) Aggregated depression level trends of U.S., NY, CA, and FL from Mar 3rd, 2020 to May 22nd, 2020. (d) Top 5 topics (state-level) after the announcement of the U.S. National Emergency.

R3: Depression and Self-Harm Risk Assessment in Online Forums

Objectives:

1. Identify the users with depression on general forum like Reddit.
2. Estimate the risk of self-harm indicated by posts in a more specific mental-health support forum.

Previous Models:

1. CLPsych 2016 investigated approaches for detecting the self-harm risk of mental health forum posts.
2. Linear classifiers with some sort of feature engineering; a combination of sparse bag of words and dense (doc2vec) representation of the target forum posts
3. A stack of feature-rich Random Forest and linear Support Vector Machines
4. RBF SVM classifier
5. Various contextual and psycholinguistic features

Our Model:

Depression Detection:

- Shared architecture based on a CNN, a merge layer, model-specific loss functions and an output layer.

Self-harm risk assessment:

- In the case of self-harm risk assessment, ordinal nature of self-harm risk labels (i.e., green, amber, red, and crisis) can improve performance
- Categorical Cross Entropy loss with softmax activation functions for user level classification model
- MSE uses an output layer with a linear activation function and rounded the output in the interval [0, t-1]

R3: Depression and Self-Harm Risk Assessment in Online Forums

Datasets:

1. For depression dataset construction, Reddit Self-reported Depression Diagnosis (RSDD) dataset is used.
2. For self-harm risk assessment, use data from mental health forum posts from ReachOut.com.

Results:

1. The result on self-harm risk assessment, the model with categorical Cross Ent. F1 Non-green: 0.50, F1 Flagged: 0.89, Acc Flagged: 0.93, F1 Urgent: 0.70, Acc Urgent: 0.94, F1 All: 0.61, Acc: 0.89
2. With 10 fold cross validation, the increase in the highest non-green F1 from 0.50 to 0.87, suggest there may be qualitative differences between the training and testing sets.
3. The best-performing method on the test set, Categorical Cross Ent., performs the worst on the training set.
4. Similarly, the worst-performing method on the test set, MSE, performs the best on the training set.

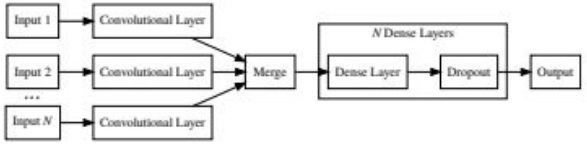


Figure 1: The general neural network architecture shared among our user and post classification models. Each input (e.g., each of a user's posts) is processed by a convolutional network and merged to create a vector representation of the user's activity. This vector representation is passed through one or more dense layers followed by an output layer that performs classification. The type of input received, merge operation, and output layer vary with the specific model.

R3: Depression and Self-Harm Risk Assessment in Online Forums

Method		Convolution			Dense Layers	Dropout	Class Balance
		Size	Filters	Pool Len.			
Reddit	Cat. Cross Ent.	3	25	all (avg)	1 w/ 50 dims	0.0	Sampled
ReachOut	Cat. Cross Ent.	3	150	3 (max)	2 w/ 250 dims	0.3	Weighted
	MSE	3	100	3 (max)	2 w/ 250 dims	0.5	Sampled
	Class Metric	3	100	3 (max)	2 w/ 150 dims	0.3	Sampled

Table 1: The hyperparameters used by each model.

Method	Precision	Recall	F1
BoW - MNB	0.44	0.31	0.36
BoW - SVM	**0.72**	0.29	0.42
Feature-rich - MNB	0.69	0.32	0.44
Feature-rich - SVM	0.71	0.31	0.44
User model - CNN	0.59	**0.45**	**0.51**

Table 2: Performance identifying depressed users on the Reddit test set. The differences between the CNN and baselines are statistically significant (McNemar's test, $p < 0.05$).

Method	Non-green	Flagged		Urgent		All	
	F1	F1	Acc.	F1	Acc.	F1	Acc.
Baseline (Milne et al., 2016)	0.31	0.78	0.86	0.38	0.89	-	-
Kim et al. (2016)	0.42	0.85	0.91	0.62	0.91	0.55	0.85
Malmasi et al. (2016)	0.42	0.87	0.91	0.64	0.93	0.55	0.83
Brew (2016)	0.42	0.78	0.85	0.69	0.93	0.54	0.79
Cohan et al. (2016)	0.41	0.81	0.87	0.67	0.92	0.53	0.80
Categorical Cross Ent.	**0.50**	**0.89**	**0.93**	0.70	**0.94**	**0.61**	**0.89**
MSE	0.42	0.80	0.85	0.64	0.93	0.53	0.78
Class Metric	0.46	0.79	0.84	0.70	**0.94**	0.56	0.80
Class Metric (Ordinal)	0.47	0.88	**0.93**	**0.72**	0.93	0.59	0.87

Table 3: Self-harm risk assessment performance on the ReachOut test posts. F1 and accuracy are aggregated as specified by CLPsych '16. The reported results for the other methods are the official numbers from (Milne et al., 2016). The differences in performance between the following method pairs are statistically significant (McNemar's test, $p < 0.05$): *Categorical Cross Ent.* and *Class Metric*, *MSE* and *Categorical Cross Ent.*, *MSE* and *Class Metric (Ordinal)*, and *Class Metric (Ordinal)* and *Class Metric*.

Method	Non-green	Flagged		Urgent		All	
	F1	F1	Acc.	F1	Acc.	F1	Acc.
Categorical Cross Ent.	0.54	0.87	0.89	0.69	0.91	0.63	0.80
MSE	**0.87**	**0.95**	**0.96**	**0.91**	**0.98**	**0.89**	**0.93**
Class Metric	0.73	0.90	0.91	0.81	0.94	0.78	0.86
Class Metric (Ordinal)	0.85	**0.95**	**0.96**	0.89	0.97	0.88	0.92

Table 4: Self-harm risk assessment performance on the ReachOut training set using 10-fold cross validation. *Categorical Cross Ent.* performs substantially worse than on the test set, while *MSE* performs substantially better. *Class Metric (Ordinal)* continues to perform well. The difference in performance between the following method pairs are statistically significant (McNemar's test, $p < 0.05$): *Categorical Cross Ent.* and *MSE*, *Categorical Cross Ent.* and *Class Metric*, *Categorical Cross Ent.* and *Class Metric (Ordinal)*, *MSE* and *Class Metric*, and *Class Metric* and *Class Metric (Ordinal)*.