Emotion Analysis for Detecting Signs of Mental Health Issues from Social Media

Diana Inkpen

Aug 20, 2018

uOttawa

L'Université canadienne
Canada's university

Université d'Ottawa | University of Ottawa

My team

Ruba Skaik
PhD student

Prasadith Buddhitha
PhD student

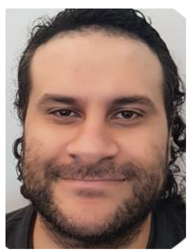
Romualdo Alves Pereira
PDF

Zunaira Jamil
MSc

Claudie Loranger
Phycologist

Kenton White
Advanced Symbolics

Ahmed Orabi
PDF



Mahmood Orabi
PDF

Shy Huang
PhD student

Dooa Ibrahim
PhD student

David van Bruwaene
VISR / SafeToNet

Osmar Zaiane
University of Alberta

Sandra Bringay
Université de Montpelier

uOttawa

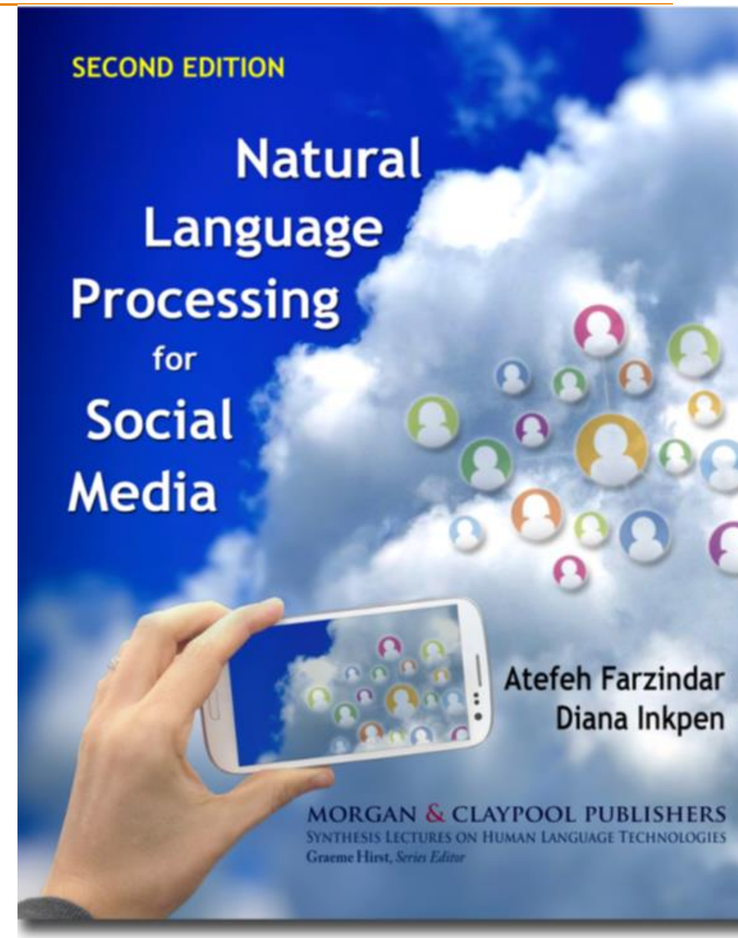
Outline

- Social Media Applications
- Mental Health
- Child Safety
- Datasets
- Methods
- Results
- Application Scenarios
- Conclusion and Future Work

uOttawa

Applications of social media (NLP4SM Book, Ch 4)

- Health care applications
- Financial applications
- Predicting voting intentions
- Security and defence applications
- Disaster response applications
- NLP-based user modelling
- Applications for entertainment
- Media monitoring



uOttawa

Health care applications

- Many online platforms where people discuss their health:
  - specialized forums, for various topics. The language is often informal and medical terms can be found, but most of the language is lay. Various kinds of information can be extracted automatically from such postings and discussions.
  - **Opinions and arguments pro and cons** topics such as: vaccinations, **mammographies**, **new born genetic screening**.
- Need privacy protection: detection of personal health information (PHI) such as names, dates of birth, addresses, health insurance numbers.
- **Detection early signs of mental heath problems (depression, suicidal ideation, etc.).**

NLP-based user modelling

- Learn user profiles based on their social media behaviour (all the postings of a user).
- **Modelling user’s personality.**
  - ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media and the hared tasks on Computational Personality Recognition 2014 and 2013.
  - Big Five model: extraversion, emotional stability, agreeableness, conscientiousness and openness to experience.
- **Modelling user’s health profile.**
- Modelling gender and ethnicity. Nationality. Race.
- Modelling user’s political orientation.
- Modelling user's life events.
- **Modelling user’s location.**

uOttawa

Related work

- **Social media and self disclosure**
  - high self-disclosure
- **Use of social media platforms to identify mental disorders**
  - surveys with lower response rates
  - interviews/surveys vulnerable to memory bias
  - social media offers a natural setting
  - linguistic and behavioural attributes (e.g., use of first person singular pronouns, **emotion**, social activities , network relationships)

- Cyberbullying detection (classifiers)
- Substance abuse (statistics)

uOttawa

Related work

- **Detecting mental disorders**
  - Detecting insomnia and distress
    - Insomnia: pronouns, verbs, auxiliary verbs, **higher negative affect, lowered positive affect, sadness, anger, anxiousness** and use of present tense words (LIWC categories)
  - Detecting postpartum depression
    - level of activity, **type of emotion and its level of intensity**, dominance and its effect, linguistic style markers (e.g., articles, auxiliary verbs), number of replies, etc. (LIWC)
  - Detecting depression and its level
    - engagement, egocentric social graph, depression language, **emotion** and linguistic style, time of post, etc. (LIWC, topic modeling)
  - Detecting Post Traumatic Stress Disorder (PTSD)
    - unigram and character n-gram language models

uOttawa

Related work

Detecting mental disorders

- Detecting Attention Deficit Hyperactivity Disorder (ADHD), Generalized Anxiety Disorder, Bipolar Disorder, Borderline Personality Disorder, Depression, Eating Disorders (anorexia, bulimia), obsessive compulsive disorder, Post Traumatic Stress Disorder (PTSD), schizophrenia, and seasonal affective disorder.
  - time, sentiment, exercise activities, etc. (language models, open-vocabulary, character n-grams, LIWC)
- CLPsych 2015 shared task
  - Identify PTSD from the control group, depression from the control group and depression from PTSD.
  - Features derived using supervised LDA, supervised anchors (for topic modeling), lexical TF-IDF, and combinations.
- CLPsych 2016 and 2017 shared task
  - Automatically prioritise content in online peer-support forum ReachOut.com by how urgently it requires moderator attention.

  Distress level: 0 (green), 1 (amber), 2(red), 4 (**crisis**).

uOttawa

Related work

- **Detecting suicide ideation**
  - the Werther effect
    - affective attributes (positive, negative), cognitive attributes, linguistic style, social attributes, etc.
    - published after a suicide: negativity, greater cognitive biases, lower lexical density, less concern about personal and social aspects, less concern about the future, more use of first person singular pronouns, more self attention, posts are longer with greater self-disclosure, etc.
  - key markers on suicidal ideation and behaviour: “want to die” vs. “want to commit suicide”
  - distinguish suicide ideation from report of suicide, suicide awareness posts and references made to suicide
  - shifts to suicidal ideation from mental health forums
  - topic models to identify suicide ideation

uOttawa

Our tasks

- **Detecting signs of mental health issues** (**depression**, **self-harm and suicide ideation**, **distress level**)
- **Children behaviour** (**aggression, sexting, cyberbullying, substance abuse**)
- **User modelling**: gender, age, location**,** **personality**, etc.

uOttawa

Datasets

- Depression
  - **Bell Let's Talk** (tweets) (ours)
  - CLPsych 2015 shared task dataset (tweets)
  - Georgetown dataset (reddit forum data)

- CLPsych 2017 shared task dataset ReachOut.com forum posts labelled with **distress level**

- Cyberbullying dataset (CB)
- Shared task on aggression identification dataset at TRAC 2018 (Facebook posts, tweets)

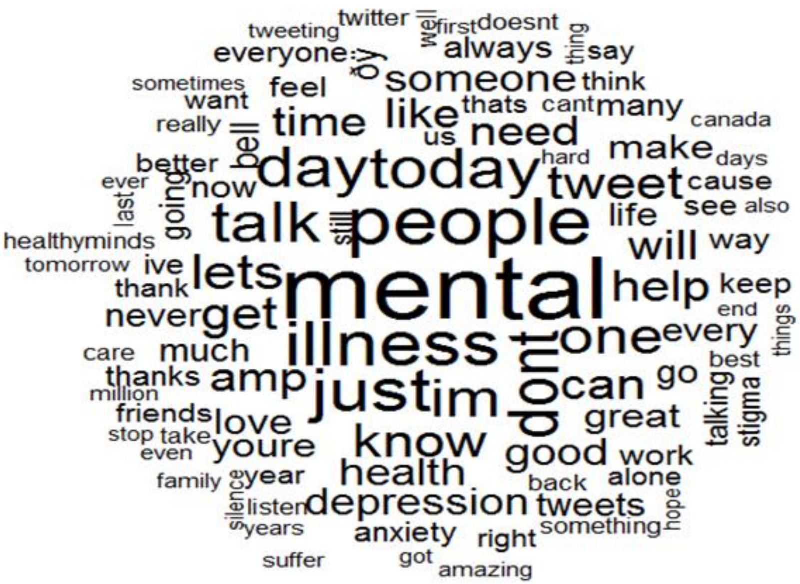
- **VISR/Safe2Net dataset** (multi-label, multi-platform) (ours)

uOttawa

Bell Let's Talk dataset

We collected all the #BellLetsTalk tweets from 2015. 156,612 public tweets were obtained from 25,362 users.
We filtered out fundraisings tweets.
We selected the tweets of 160 users to be checked by 2 annotators (k=0.67).

	Depression	Control
Number of users	69	91
Number of labeled tweets (from 60 users)	820	7,993



Examples of Bell Let's Talk tweets

Tweet 1 **"need change suffering shouldnt wait treatment "**

Tweet 2 **" Talking can save ur life"**

uOttawa

CLPsych 2015 dataset

	Control	Depression	PTSD
Number of users	572	327	264
Number of tweets in each category (not labelled)	1,250,606	742,793	544,815
Average age	24.4	21.7	27.9
Gender (female) distribution per class	74%	80%	67%

uOttawa

VISR dataset

Old version:

The cyberbullying part of the dataset: 14,193 online posts.

- There are 1,753 instances labeled as positive for cyberbullying, and the rest 12,440 instances negative.
- 3 annotators, agreement 95%, k=0.805
- Initial "real issues": 289 cyberbullying 2,983 non-bullying
- Also labels for cyber-aggression and reported bullying.

New version of the dataset is in progress with multiple labels.

- 28,523 posts
- 7 categories: **aggression, anxiety, depression, distress, sexuality, substance use, violence**

Cyberbullying dataset

- What is considered as cyberbullying:
  - Offensive
  - Harmful
  - Repeated
  - Anytime
  - **Anywhere**

- The CB dataset (Huang et al., 2014) consists of 2,150 pairs of users who have a number of 4,865 inter-changed messages.
- Only 91 messages were labeled as cyberbullying.

uOttawa

Shared Task on Aggression Identification 2018 dataset First Workshop on Trolling, Aggression and Cyberbullying at COLING 2018

Category	Train	Dev	Test		Total
			Facebook	Twitter	
Covertly aggressive	4,240	1,057	142	413	5,852
Overtly aggressive	2,708	711	144	361	3,924
Non-aggressive	5,051	1,233	630	483	7,397

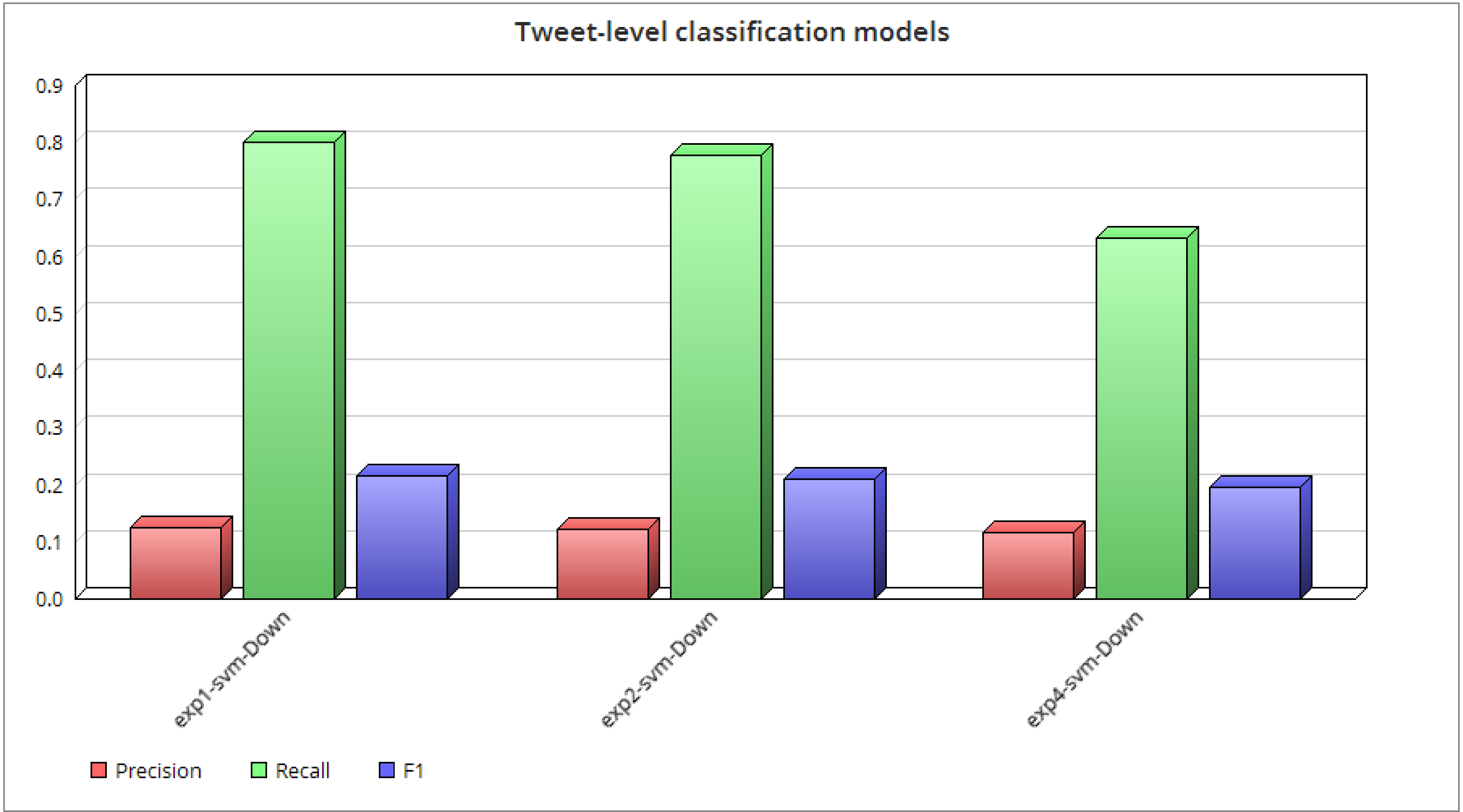
uOttawa

Methods

- Unsupervised learning – pretrained or optimized word embeddings
- Supervised classifiers (SVM with smart features, Deep Learning)
- Multi-task learning (MTL)
  - Train each model independently
  - Constrained shared layer
  - Freeze and train joint layers
  - Adaptive threshold layer
- Domain expert knowledge

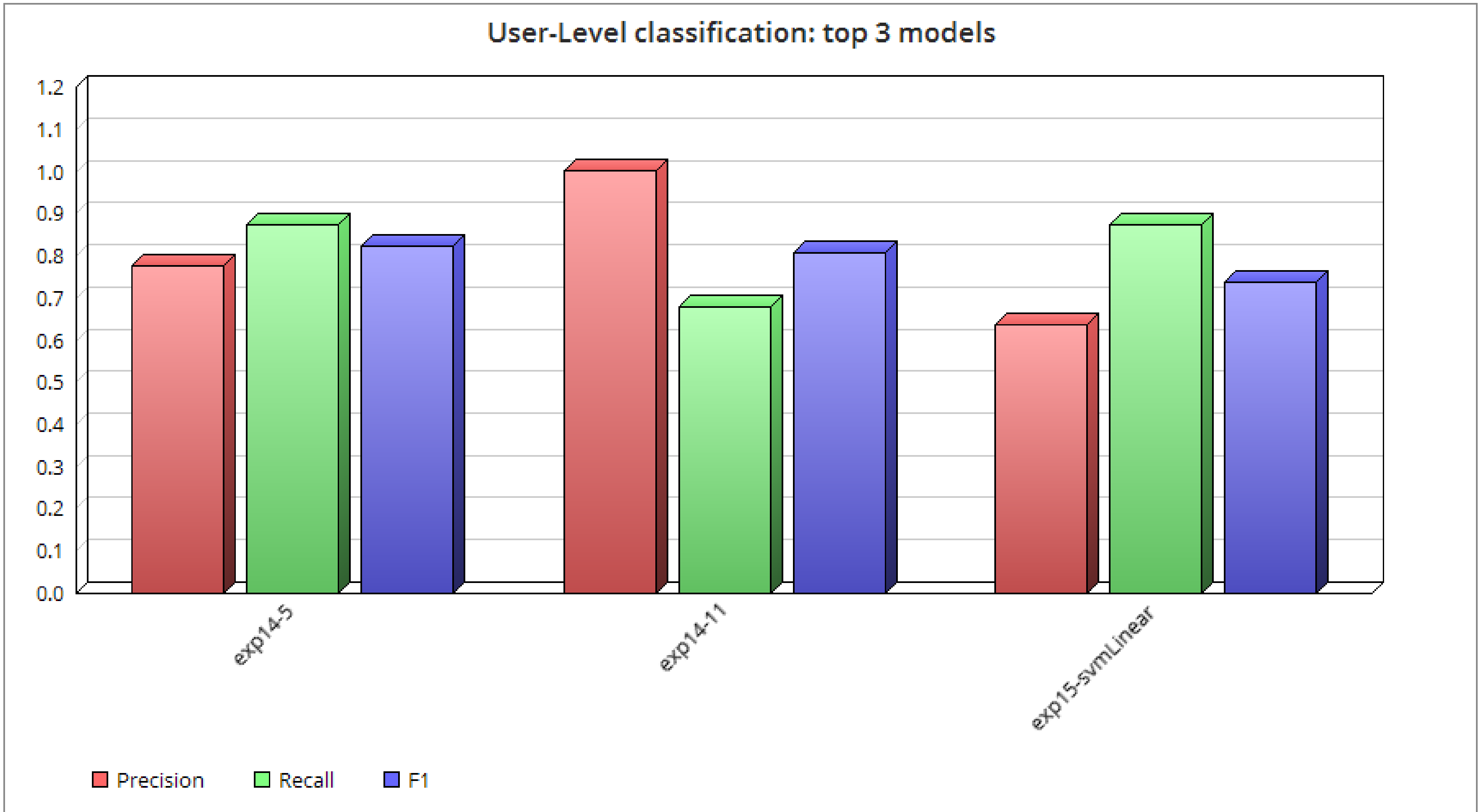
Results: Monitoring tweets for signs of depression
#BellLetsTalk **tweets** (8,753 tweets from 60 users)

SVM with various features (Jamil et al. @CLPsych 2017)

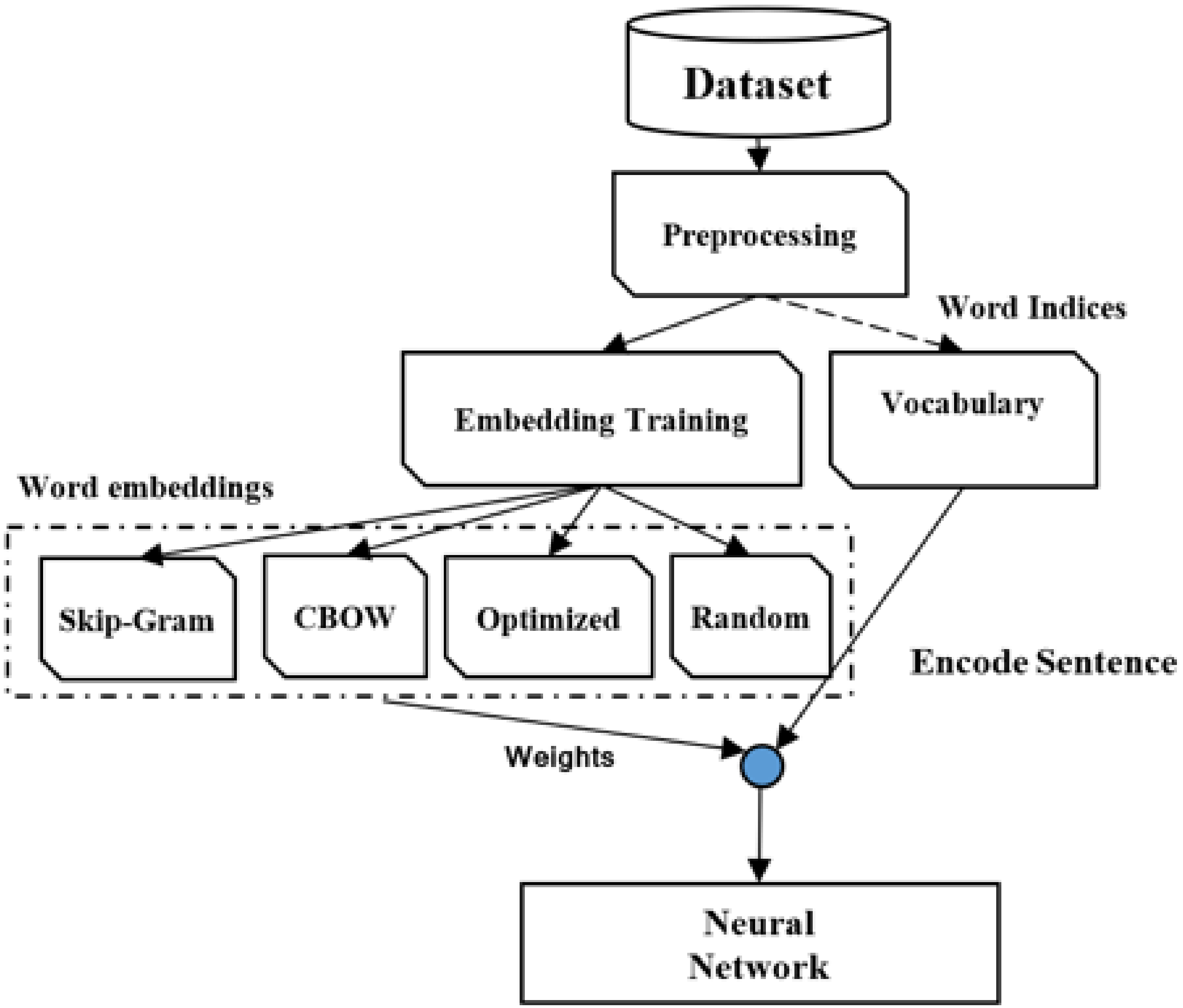


Results: Detecting at-risk **users**
#BellLetsTalk dataset (160 users)

SVM with various features (Jamil et al. @CLPsych 2017). The predictions of the tweet-level classifier were sued as features for the user-level classifier.

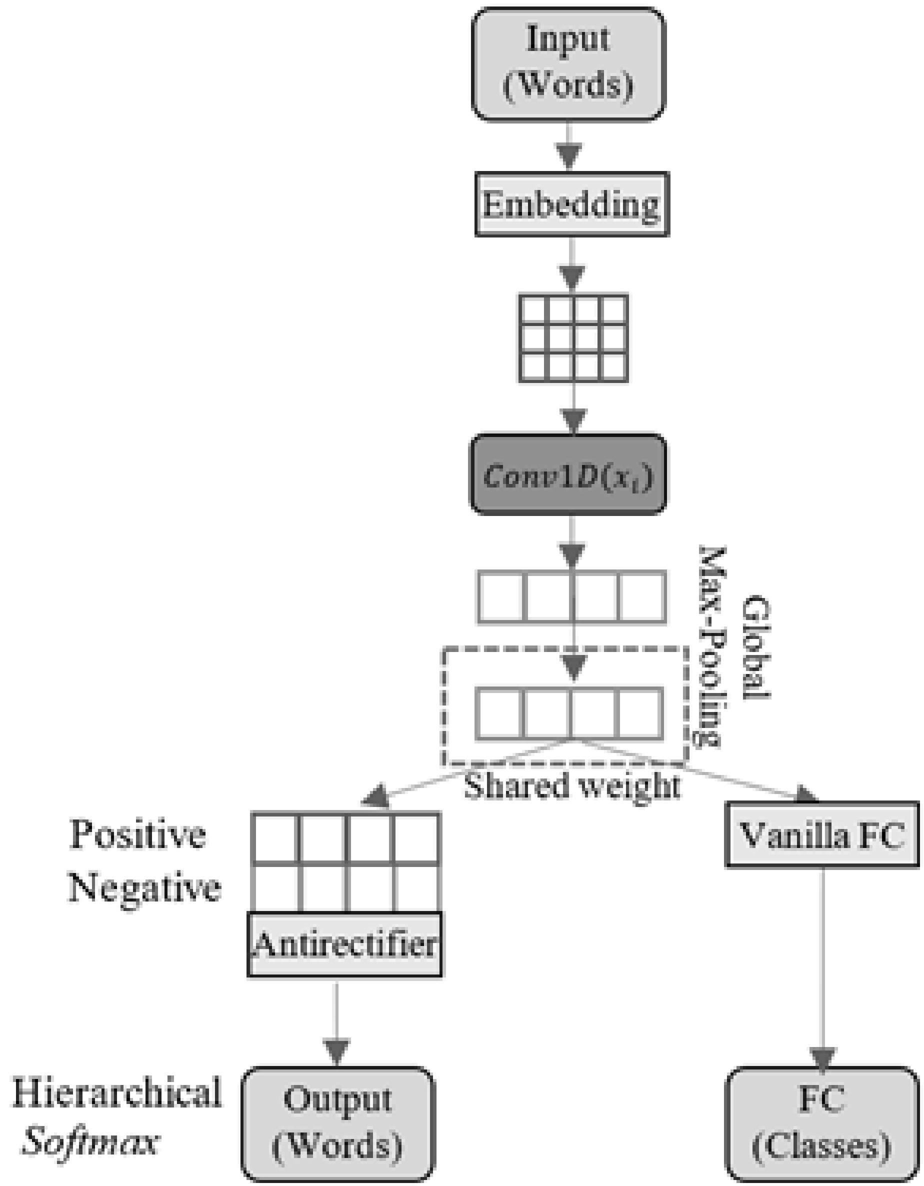


Method (Deep Learning) (Orabi et al. @CLPsych 2018)



uOttawa

Network architecture



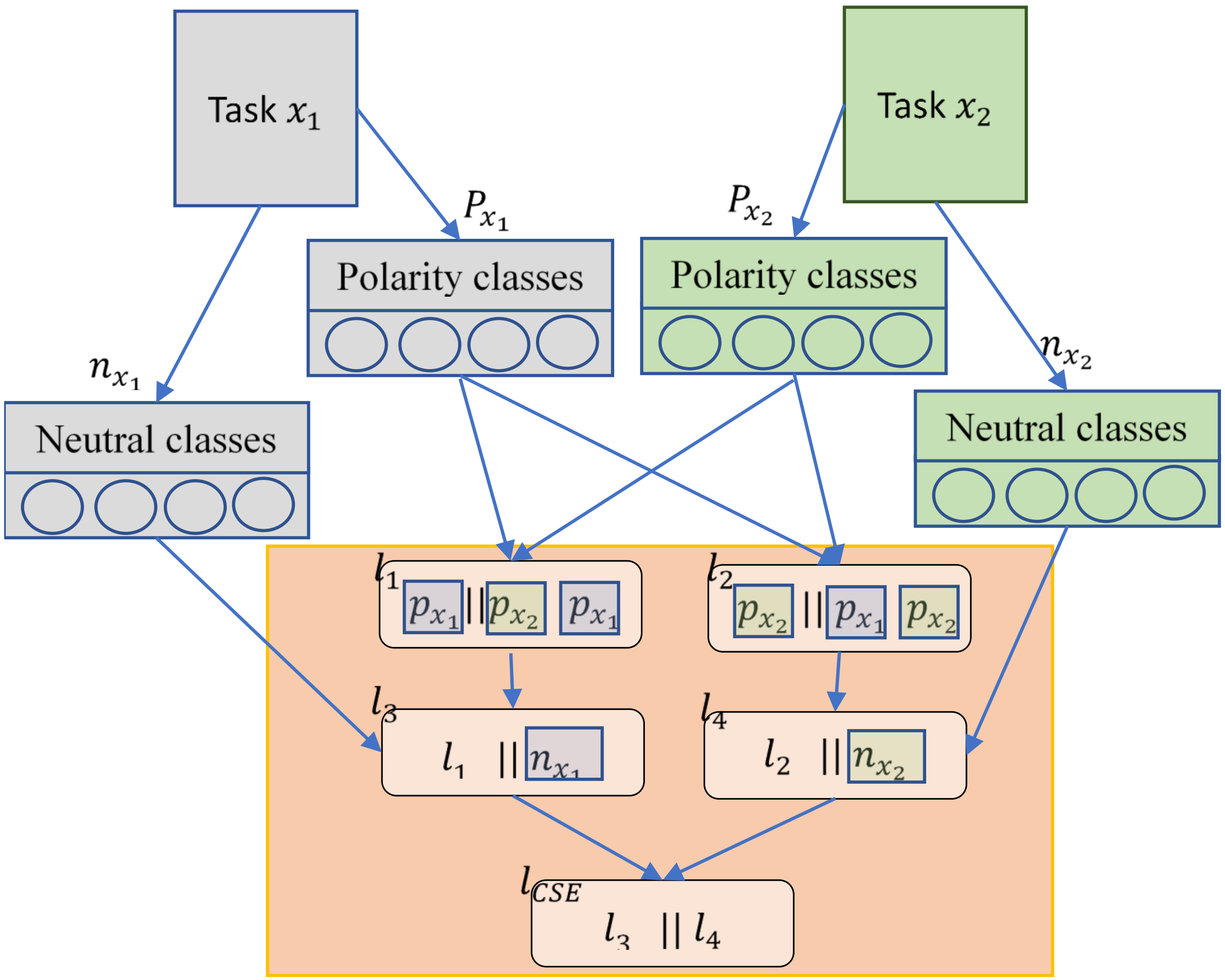
Results of DL depression classifiers on the CLPsych 2015 dataset (cross-validation)

Model	Embedding	Accuracy	Precision	Recall	F1	AUC
SVM Baseline		77.4%	0.776	0.774	0.774	0.844
CNNWithMax	CBOW	60.7%	0.380	0.542	0.430	0.544
	Skip-gram	79.8%	0.797	0.789	0.784	0.879
	Trainable	80.8%	0.804	0.820	0.801	0.909
	Optimized	87.9%	0.874	0.870	0.869	**0.951**
MultiChannelPoolingCNN	CBOW	49.6%	0.332	0.536	0.375	0.556
	Skip-gram	78.8%	0.805	0.756	0.760	0.883
	Trainable	73.6%	0.726	0.727	0.720	0.824
	Optimized	87.5%	0.872	0.866	0.864	**0.950**
MultiChannelCNN	CBOW	76.2%	76.478	0.717	0.720	0.803
	Skip-gram	81.1%	0.811	0.779	0.786	0.892
	Trainable	82.2%	82.770	0.799	0.803	0.870
	Optimized	85.6%	0.858	0.840	0.841	**0.935**
BiLSTM	Trainable	77.589	76.687	0.749	0.751	0.832
	Optimized	78.1%	76.555	0.757	0.760	0.826

Results of DL depression classifiers on Bell Let's Talk data (to test generalization)

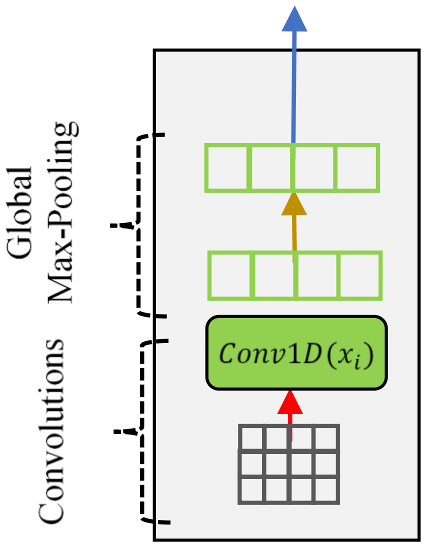
	Embedding	Accuracy	Precision	Recall	F1	AUC
SVM Baseline		73.4%	0.733	0.740	0.734	0.718
CNNWithMax	CBOW	61.6%	0.632	0.645	0.612	0.687
	Skip-gram	72.0%	0.718	0.742	0.713	0.743
	Trainable	64.9%	0.683	0.696	0.647	0.751
	Optimized	81.8%	0.805	0.834	0.809	**0.920**
MultiChannelCNN	CBOW	72.0%	0.689	0.661	0.668	0.734
	Skip-gram	62.3%	0.576	0.573	0.574	0.586
	Trainable	68.1%	0.683	0.703	0.674	0.773
	Optimized	83.1%	0.816	0.844	0.822	**0.923**
MultiChannelPoolingCNN	CBOW	51.9%	0.690	0.629	0.507	0.682
	Skip-gram	64.2%	0.693	0.700	0.642	0.752
	Trainable	60.3%	0.549	0.545	0.545	0.525
	Optimized	82.4%	0.808	0.834	0.815	**0.888**
BiLSTM	Trainable	63.6%	0.636	0.651	0.627	0.733
	Optimized	80.5%	0.805	0.838	0.800	**0.914**

Method: Multi-Task Learning

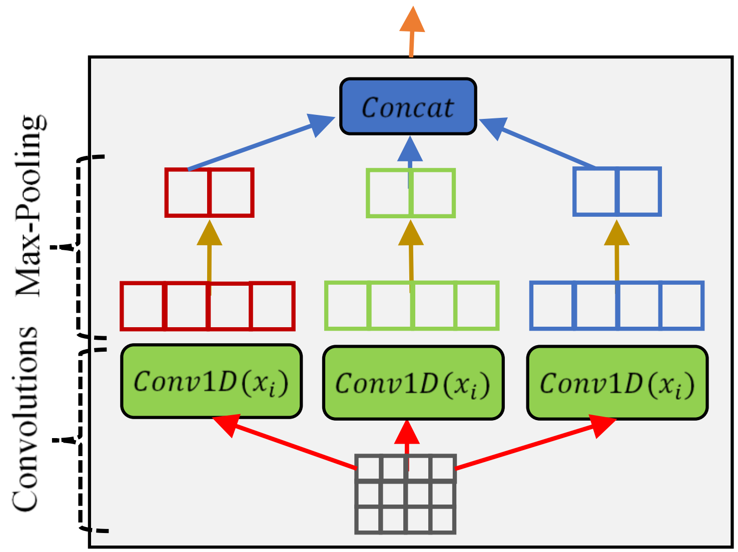


uOttawa

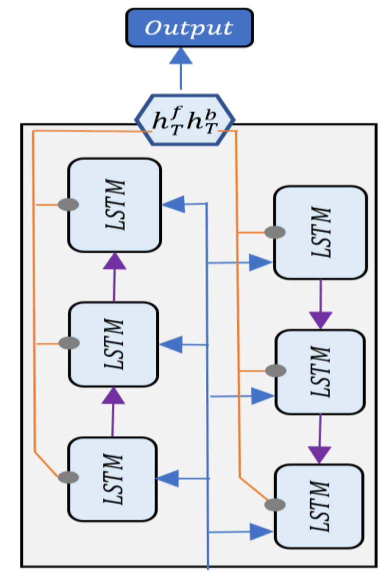
Neural network models



(a) CNNGlobal



(b) MultiCNNPooling



(c) Bidirectional LSTM

uOttawa

Results: VISR Cyber-Bullying Dataset (cross validation)

Message level

SVM baseline AUC: $\boxed{0.615}$

Model	Parameter		Embedding	Accuracy	Precision	Recall	F1	AUC
BiLSTM	Pretrained	Trainable	Word2Vec	82.7%	0.885	0.827	0.847	0.856
			GloVe	83.4%	0.894	0.834	0.854	0.877
		Frozen	Word2Vec	78.9%	0.878	0.789	0.815	0.848
			Glove	80.7%	0.897	0.807	0.835	0.887
	Trainable			80.6%	0.880	0.806	0.831	0.842
CNNGlobal	Pretrained	Static	Word2Vec	83.7%	0.892	0.837	0.855	0.885
			GloVe	87.2%	0.897	0.862	0.874	0.897
		Frozen	Word2Vec	87.9%	0.898	0.869	0.879	0.901
			GloVe	84.3%	0.896	0.843	0.861	0.896
	Trainable			81.4%	0.884	0.814	0.837	0.858
MultiCNN	Pretrained	Static	Word2Vec	82.3%	0.889	0.823	0.845	0.880
			GloVe	82.9%	0.898	0.829	0.851	0.900
		Frozen	Word2Vec	85.7%	0.899	0.857	0.871	**0.912**
			GloVe	89.8%	0.896	0.830	0.851	0.830
	Trainable			82.9%	0.883	0.829	0.847	0.858

Results: Emotion MTL Task

S refers to a single task, while M refers to training with MTL approach

	Embedding	Acc. Avg		Precision Avg		Recall Avg		F1 Avg.		AUC	
		S	M	S	M	S	M	S	M	S	M
BiLSTM	Word2Vec	74.2%	79.3%	0.876	0.896	0.742	0.781	0.785	0.822	0.815	0.873
	GloVe	74.7%	77.3%	0.868	0.900	0.747	0.773	0.789	0.815	0.789	0.856
	Trainable	71.1%	72.2%	0.859	0.862	0.711	0.721	0.761	0.770	**0.708**	**0.803**
CNNGlob al	Word2Vec	80.2%	83.7%	0.861	0.906	0.802	0.837	0.824	0.860	0.727	0.889
	GloVe	73.8%	77.7%	0.858	0.901	0.738	0.777	0.781	0.814	0.715	0.857
	Trainable	81.8%	91.3%	0.862	0.944	0.818	0.913	0.837	0.924	**0.728**	**0.953**
MultiCNN	Word2Vec	80.5%	80.4%	0.864	0.904	0.805	0.804	0.826	0.837	0.738	0.880
	GloVe	79.3%	72.0%	0.857	0.889	0.793	0.720	0.818	0.768	0.711	0.809
	Trainable	82.4%	90.2%	0.862	0.942	0.824	0.902	0.840	0.915	**0.728**	**0.956**

Joined 3 datasets, CrowdFlower text emotion (https://www.crowdflower.com/wp-content/uploads/2016/07/text_emotion.csv) , blogs (Aman & Szpakowicz, 2015), and tweets (Buechel & Hahn, 2017).

8,638 neutral, 16,252 joy, 10,290 sadness, 5,219 anger, 11,971 fear, 4,601 trust, 2,398 disgust, 6,196 surprise, and 1,526 anticipation instances.

uOttawa

Results - Cyberbullying MTL Task on CB dataset (to test generalization)

	Parameters		Embedding	Accuracy		Precision		Recall		F1		AUC	
				S	M	S	M	S	M	S	M	S	M
BiLSTM	Pretrained	Static	Word2Vec	84.6%	92.0%	0.972	0.971	0.846	0.920	0.901	0.943	0.740	0.749
			GloVe	95.2%	95.1%	0.971	0.972	0.952	0.951	0.961	0.961	0.798	0.798
		Frozen	Word2Vec	95.3%	86.9%	0.970	0.972	0.953	0.869	0.961	0.914	0.809	0.815
			GloVe	95.3%	94.7%	0.971	0.972	0.953	0.947	0.961	0.958	0.826	0.829
	Trainable			79.5%	89.8%	0.972	0.972	0.795	0.898	0.870	0.931	0.731	0.739
CNNGlobal	Pretrained	Static	Word2Vec	93.5%	92.1%	0.969	0.969	0.935	0.921	0.951	0.943	0.746	0.756
			GloVe	91.3%	94.1%	0.972	0.971	0.913	0.941	0.939	0.955	0.799	0.761
		Frozen	Word2Vec	96.9%	94.5%	0.971	0.972	0.969	0.945	0.970	0.957	0.816	0.821
			GloVe	86.6%	91.1%	0.972	0.972	0.866	0.911	0.913	0.938	0.809	0.809
	Trainable			86.7%	91.3%	0.971	0.972	0.867	0.913	0.913	0.939	0.733	0.799
MultiCNN	Pretrained	Static	Word2Vec	91.9%	92.8%	0.970	0.970	0.919	0.928	0.942	0.947	0.746	0.742
			GloVe	92.5%	94.5%	0.972	0.972	0.925	0.945	0.946	0.957	0.809	0.830
		Frozen	Word2Vec	93.8%	95.1%	0.973	0.972	0.938	0.951	0.953	0.960	0.821	0.831
			GloVe	91.3%	88.7%	0.973	0.974	0.913	0.887	0.940	0.925	$\boxed{\textbf{0.806}}$	$\boxed{\textbf{0.834}}$
	Trainable			91.9%	86.9%	0.971	0.971	0.919	0.869	0.942	0.914	0.750	0.755

SVM baseline: AUC: $\boxed{0.600}$ Adding SN features 0.750

Results - Aggression classification on the TRAC 2018 shared task test set

Dataset		F1 (weighted)
Twitter	Baseline	34.77%
	Our model	56.90%
Facebook	Baseline	35.35%
	Our model	59.74%

Main task: classification into 3 classes (covertly aggressive, overtly aggressive, and non-aggressive)

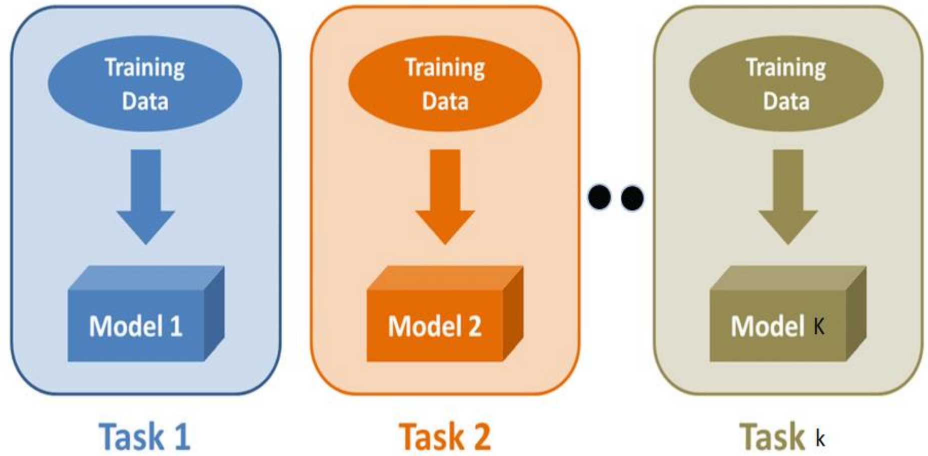
Method: MTL aggression and **emotion** classification

uOttawa

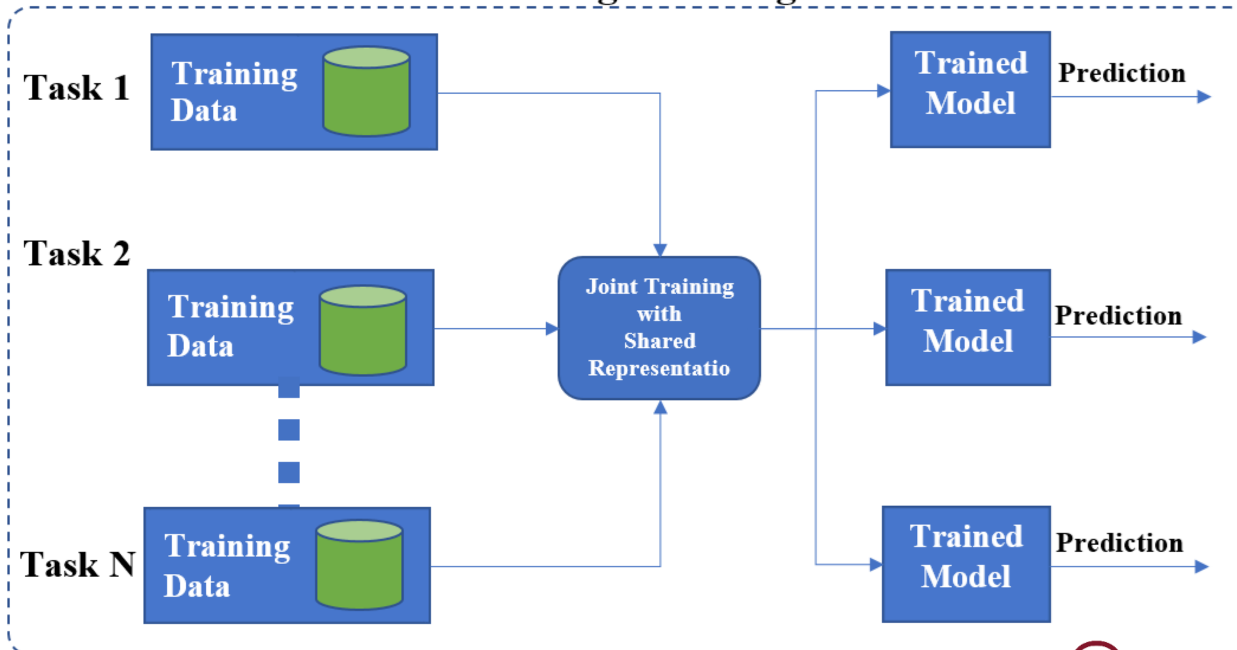
MTL: more than two tasks

Approach:

- Train for each independently.
- Constrain shared layer.
- Freeze and train joint layers.
- Adaptive threshold output layer.



Multitasking Learning



uOttawa

MTL results for 10 tasks

Task	Accuracy	Precision	Recall	F1	AUC
aggression	89.6%	0.548	0.619	0.581	0.913
bullying	86.8%	0.601	0.671	0.617	0.908
sexuality	95.8%	0.729	0.729	0.729	0.950
sentiment	80.0%	0.765	0.867	0.812	0.890
mood	75.1%	0.707	0.856	0.775	0.840
joy-sadness	80.1%	0.784	0.823	0.803	0.876
anger-fear	81.5%	0.755	0.906	0.823	0.892
surprise-anticipation	70.4%	0.673	0.833	0.744	0.752
trust-disgust	86.0%	0.839	0.918	0.876	0.928
mental-health	83.4%	0.778	0.912	0.840	0.902

uOttawa

Application scenarios

- Message-level models
  - for depression, etc.
  - for the child safety app
- User-level models
  - for a example a psychologist can monitor patients with their consent, or post-monitor patients who finished therapy to get alerts about relapses
- Population-level models
  - for better distribution of health care spending

Population-level predictions

- Many Canadians believe that Aboriginal youth and youth in Northern Communities are at higher risk of suicide then the general population. Let's assume that is true.
- Previous research showed that shame, guilt, and somatic complaints are correlated with suicide ideation.
- We plan to test this **hypothesis** with an experiment. Create a Twitter stream for users from Northern Communities (communities in Northern Ontario that had a cluster of suicides, or Aboriginal communities, like Nunavut). Create a separate Twitter stream for a white affluent community like Edmonton or Calgary.
- Then **compare mentions of shame, guilt, and somatic complaints** from each community. Are the numbers statistically different?

- Train **a user-level classifier for suicide ideation and attempts**. Apply on the above data. See if there are correlations.

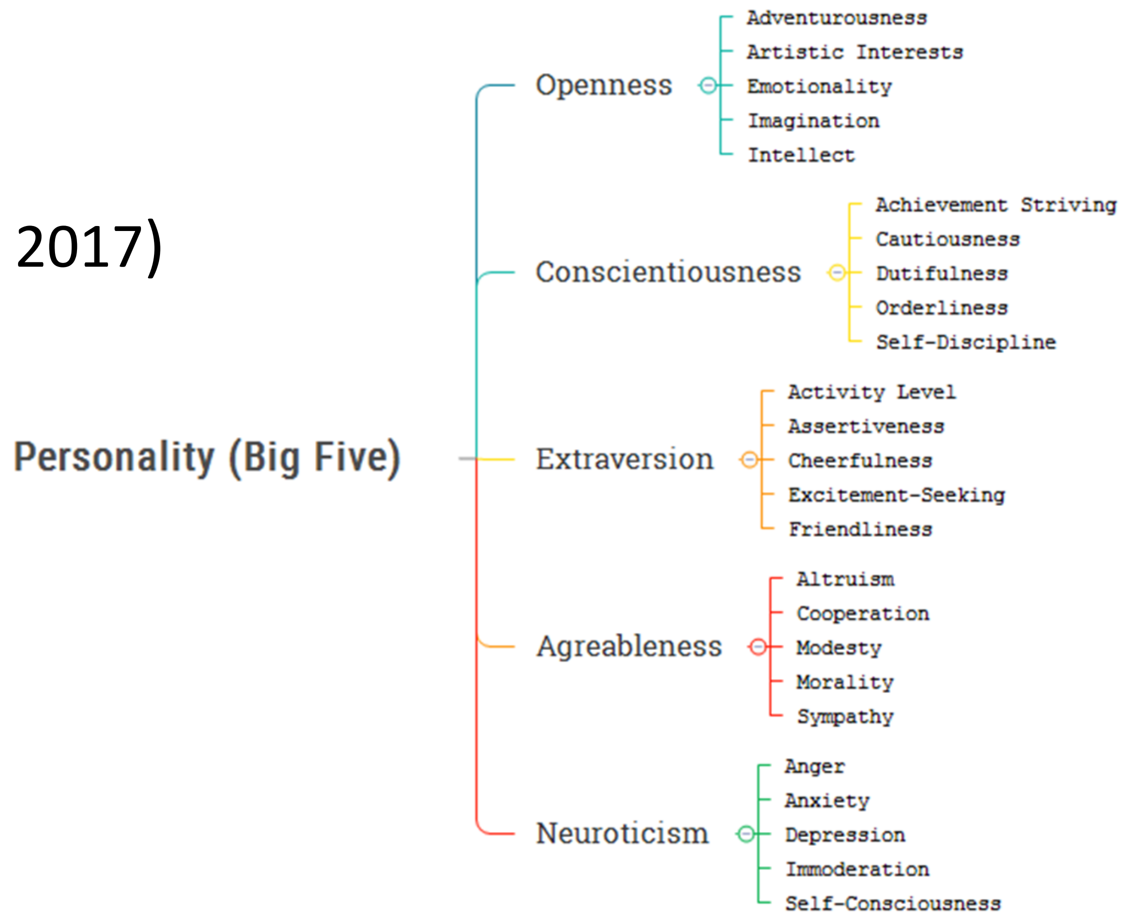
uOttawa

User Modelling

- Gender and age detection (PAN @CLEF 2015 dataset)

- Personality detection

  (Alves Pereira and Inkpen, 2017)

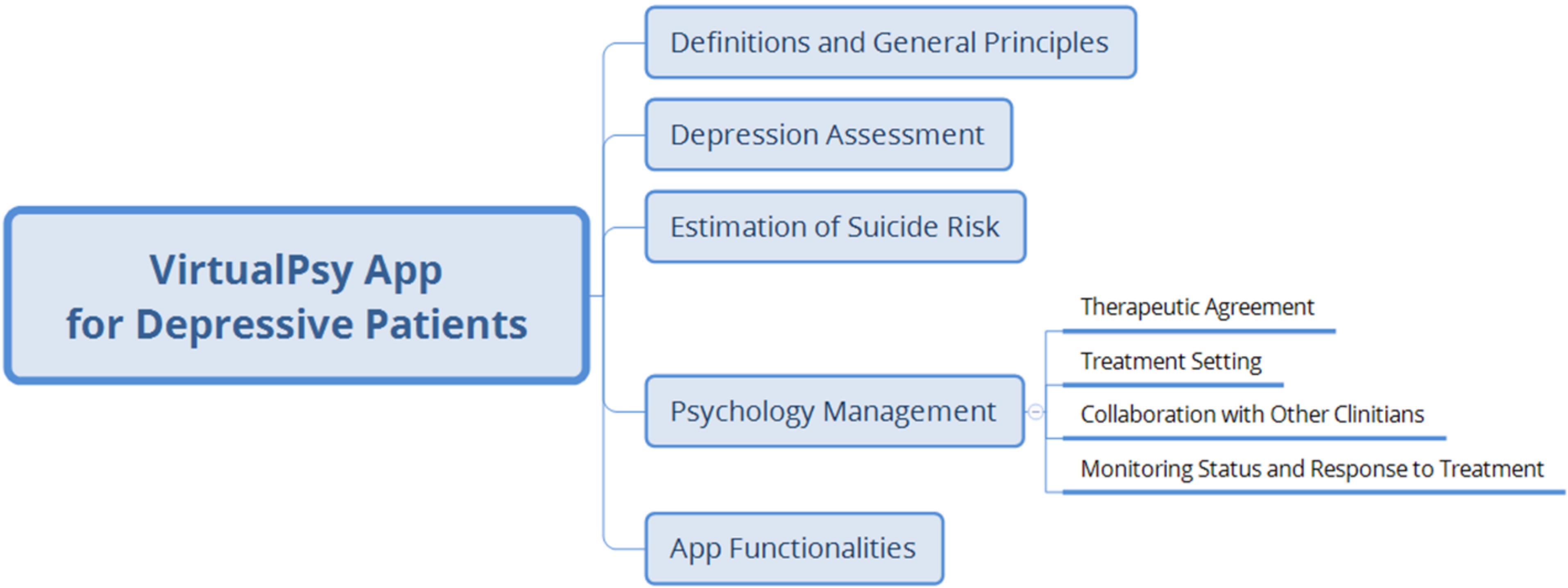


- Location detection

  (Liu and Inkpen, 2017)

uOttawa

Tool for Post-Monitoring Depression



Tool for post-monitoring depression

VirtualPsy

Essential features (PHQ-9)

- Each item is a symptom of depression that would have to be identified by the application. When 5 or more symptoms are present most days for more than two weeks, it indicates clinical depression.
- 1-2 symptoms: Low risk
- 3-4 symptoms: Moderate risk
- 5 symptoms or more: High risk

Emotions

- All **negative emotions** (sadness, anger, anxiety, guilt) should be identified by the application and considered when identifying depression.
- Depression = Low positive emotions + High negative emotions

uOttawa

Tool for post-monitoring depression

VirtualPsy

<u>Additional features</u>

- Suicidal ideation = low emotional stability + low extraversion + low agreeableness
- Depression = high neuroticism + low extraversion + low emotional stability + low conscientiousness

<u>Adverse life events</u>

- Is the person going through a difficult situation (break-up, grief, illness, conflicts…)?

<u>Prior knowledge of the user</u>

- Allow the user to insert personal information in the application in order to personalize the algorithm with known risk factors.

uOttawa

Child Safety App: SafeToNet

SafeToNet
informed safeguarding

- Parent's app & Child's app
- Parents cannot see the actual messages.
- Warnings for sexting, aggressive behaviour, offensive language, etc.
- Extensions in progress:
  - multiple languages
  - image and video processing (for pornography, bullying, etc.)

uOttawa

Ethics considerations

- Projects need ethics approval (secondary use of data).
- Public data or permission.
- Securely store the data.
- Anonymize data.
- Never identify or contact users.
- Care when application scenarios proofs of concept become applications.

uOttawa

Future work

- Multi task learning for emotion, depression and suicide ideation.
- Investigate correlations to personality traits.
- Investigate correlations to substance abuse.
- Other mental health issues (PTSD, bipolar disorder, schizophrenia, etc.)
- Extend child safety app – multi-lingual and multi-modal.
- Test the application scenarios.
- More applications scenarios
  - Monitor inmate messages for dangers to themselves or others.
  - Self monitoring for well-being.

uOttawa

References

NLP for Social Media (2017)
by Atefeh Farzindar and Diana Inkpen
Synthesis Lectures on Human Language Technologies, Morgan and Claypool, **second edition**

Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. Monitoring Tweets for Depression to Detect At-risk Users. In Proceedings of **CLPsych 2017** at ACL 2017, Vancouver, BC, Canada, August 2017, pp. 32-40.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, Diana Inkpen. Deep Learning for Depression Detection of Twitter Users. In Proceedings of **CLPsych 2018** at NAACL 2018, New Orleans, LA, USA, June 2018.

uOttawa

More references

- Romualdo Alves Pereira Jr. and Diana Inkpen. Using Cognitive Computing to Get Insights on Personality Traits from Twitter Messages. In Proceedings of the 30th Canadian Conference on Artificial Intelligence (AI 2017), Edmonton, AB, Canada, May 2017, pp. 279-283.
- Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Hierarchical versus flat classification of emotions in text. In Proceedings of the NAACL HLT 2010 Workshop on Computational approaches to analysis and generation of emotion in text, pages 140–146, Los Angeles, CA, USA, June 2010.
- Diana Inkpen, Ji Liu, Atefeh Farzindar, Farzaneh Kazemi, and Diman Ghazi. Location detection and disambiguation from Twitter messages. In Proceedings of CICLing 2015, LNCS 9042, pages 321–332, Cairo, Egypt, 2015.
- Ji Liu and Diana Inkpen. Estimating user locations on social media: A deep learning approach. In Proceedings of the NAACL 2015 Workshop on Vector Space Modeling for NLP, Denver, Colorado, USA, 2015.

uOttawa

Questions?

uOttawa