Multi-Task, Multi-Channel, Multi-Input Learning for Mental Illness Detection using Social Media Text

Prasadith Buddhitha and **Diana Inkpen**
School of Electrical Engineering and Computer Science
University of Ottawa, Canada
{pkiri056, diana.inkpen}@uottawa.ca

Abstract

We investigate the impact of using emotional patterns identified by the clinical practitioners and computational linguists to enhance the prediction capabilities of a mental illness detection (in our case depression and post-traumatic stress disorder) model built using a deep neural network architecture. Over the years, deep learning methods have been successfully used in natural language processing tasks, including a few in the domain of mental illness and suicide ideation detection. We illustrate the effectiveness of using multi-task learning with a multi-channel convolutional neural network as the shared representation and use additional inputs identified by researchers as indicatives in detecting mental disorders to enhance the model predictability. Given the limited amount of unstructured data available for training, we managed to obtain a task-specific AUC higher than 0.90. In comparison to methods such as multi-class classification, we identified multi-task learning with multi-channel convolution neural network and multiple-inputs to be effective in detecting mental disorders.

1 Introduction

Social media platforms have revolutionized the way people interact as a society and have become an integral part of everyday life where many people have started sharing their day to day activities on these platforms. Such real-time data portraying one's daily life could reveal invaluable insights into one's cognition, emotion, and behavioral aspects. With its rapid growth among different demographics and being a source enriched with valuable information, social media can be a significant contributor to the process of mental disorder and suicide ideation detection.

In the domain of mental illness detection and especially when using social media text, lack of sufficiently-large annotated data and the inability to extract explanation on the derived outcome have restricted researchers to use traditional machine learning algorithms other than state-of-the-art methods such as deep neural networks. The proposed research explores the feasibility of applying the state-of-the-art processes in combination with features identified using manual feature engineering methods to enhance the prediction accuracy while maintaining low false negative and false positive rates. Also, this research looks into detecting multiple mental disorders at the same time by sharing lower level features among the different tasks. Intuitively, learning multiple mental disorders using a single neural network architecture in comparison to using a single network to identify only one mental illness is a logical approach considering the psychological and linguistic characteristics shared among the individuals susceptible of being diagnosed with different mental disorders.

Mental illness detection in social media using Natural Language Processing (NLP) methods is considered as a difficult task due to the complex nature of mental disorders. According to the American Psychiatric Association (2013), a mental disorder is a "syndrome characterized by a clinically significant disturbance in an individual's cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning". Mental disorders have become and continue to be a global health problem. More than 300 million people from varied demographics suffer from depression (World Health Organization, 2018a), and have broader implications where 23% deaths in the world were caused due to mental and substance use disorders (World Health Organization, 2014). In Canada, one in every five Canadians has experienced some form of mental ill-

Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pages 54–64

https://doi.org/10.18653/v1/D19-62

ness (Canadian Mental Health Association, 2016). The adverse impact of mental disorders is prominent when looking into the number of Canadians who have committed suicide, where 90% of them were suffering from some form of a mental disorder (Mental Health Commission of Canada, 2016).

When considering the current treatment procedures for mental illnesses, the first step is to screen the individual for symptoms using questionnaires. With such an approach, the interviewee could be more vulnerable to memory bias and also could adapt to the guidelines prescribed by the assessor. Using such screening procedures might not expose the actual mental state of an individual, and hence, the prescribed treatments could be inadequate. Also, due to various socio-economic aspects, people with mental disorders have not been able to receive adequate treatments. The lack of sufficient treatments can be seen in countries with all types of income levels. For example, between 76% to 85% of the people from countries in low to middle-class income do not receive sufficient treatments for their mental illnesses, while 35% to 50% of the people from high-income countries do not receive adequate treatments (World Health Organization, 2018b). In addition to insufficient treatments, social stigma and discrimination have prohibited people from getting proper treatment and social support (World Health Organization, 2018b). Due to the constraints as mentioned above and social media becoming an integral part of everyday life for many individuals, researchers have identified the importance of using social media data as a source for ascertaining individuals susceptible of mental disorders (De Choudhury, 2013, 2014). Due to the rapid growth in social media users within different demographics (Statista, 2017), and the abundance of information that can be extracted about the users could bring invaluable insights that can be used to detect signs of mental illnesses and suicide ideation that can be challenging to obtain using structured questionnaires. Taking the factors mentioned above into consideration, we have proposed a solution that incorporates certain profound features manually engineered by researchers into a multi-task learning architecture, to enhance the model predictability to distinguish neurotypical users from users susceptible to mental disorders. We hope that our research will encourage other researchers to investigate further the possibilities of incorporating verified manually engineered features into architectures similar to the proposed one, to enhance the prediction accuracies in identifying users susceptible to mental disorders.

Key Contributions

- We demonstrate the applicability of a convolutional neural network with a multi-channel architecture on different classification tasks using unstructured and limited social media data.
- We illustrate the use of multi-task learning to predict users susceptible to depression and PTSD (Post Traumatic Stress Disorder).
- We built an emotion classifier to identify the emotion category (i.e., sad, anger, fear, joy) associated with the tweets posted by the users and used those categories as multiple inputs within the deep neural network architecture. We also explore the impact of using metadata (age and gender) as multiple inputs to enhance the model predictability.

2 Ethical Considerations

It is of greater importance to follow strict guidelines on ethical research conduct when the research data resembles vulnerable users who could be compromised. The researchers working with data that could be used to single out individuals must take adequate precautions to avoid further psychological distress. During our research, we have given thorough considerations to these ethical facets and have adopted strict guidelines to ensure the anonymity and privacy of the data. Similar to the guidelines proposed by Benton et al. (2017a), we have exercised strict hardware and software security measures. Our research does not involve any intervention and has focused mainly on the applicability of machine learning models in determining users susceptible to mental disorders.

3 Related work

As social media has become an integral part of ones' day-to-day-life, it will be insightful to identify to what extent an individual has disclosed her/his personal information and whether accurate and sufficient information is being published to determine whether or not a person has a mental disorder. Considering the Twitter platform, rather than just sharing depressed feelings, users

are more likely to self-disclose to the extent where they reveal detailed information about their treatment history (Park et al., 2013). The same level of self-disclosure can be identified in the Reddit forums (Balani and De Choudhury, 2015) and specifically by users with anonymous accounts (Pavalanathan and De Choudhury, 2015). Also, it was identified that personality traits and meta-features such as age and gender could have a positive impact on the model performances when detecting users susceptible to PTSD and depression (Preot et al., 2015). Similarly, we have also identified that age and gender as multiple inputs have positively impacted model predictability when used with multi-task learning.

Text extracted from social media platforms such as Twitter, Facebook, Reddit, and other similar forums has been successfully used in various natural language processing (NLP) tasks to identify users with different mental disorders and suicide ideation. Social media text was used to classify users with insomnia and distress (Jamison-Powell et al., 2012; Lehrman et al., 2012), postpartum depression (De Choudhury et al., 2013a,b, 2014), depression (Resnik et al., 2015a, 2013, 2015b; Schwartz et al., 2014; Tsugawa et al., 2015), Post-Traumatic Stress Disorder (Coppersmith et al., 2014a,b), schizophrenia (Loveys et al., 2017) and many other mental illnesses such as Attention Deficit Hyperactivity Disorder (ADHD), Generalized Anxiety Disorder, Bipolar Disorder, Eating Disorders and obsessive-compulsive disorder (OCD) (Coppersmith et al., 2015a).

With the advancements in neural network-based algorithms, more research has been conducted successfully in detecting mental disorders, despite the limited amount of data. Kshirsagar et al. (2017) have used recurrent neural networks with attention to detect social media posts resembling crisis. Husseini Orabi et al. (2018) demonstrated that using convolution neural network-based architectures produces better results compared to recurrent neural network-based architectures when detecting users susceptible to depression. Even though our experiments are to categorize users into three classes (i.e., control, depression, PTSD), the proposed multi-channel architecture have produced comparable results to the ones presented by Husseini Orabi et al. (2018) using binary classification to distinguish users susceptible to depression.

4 Proposed solution

The proposed solution consists of two key components. The first identifies the type of emotion expressed by each user using the model trained on the WASSA 2017 shared task dataset. The identified emotion categories are used as multiple inputs within the multi-task learning environment. The second component is the model that predicts users susceptible to PTSD or depression. When structuring the two neural network models (i.e., for emotion classification and mental illness detection), a common base architecture is used. The base architecture is a multi-channel Convolutional Neural Network (CNN) with three different kernel sizes (i.e., 1, 2, and 3). Through experiments, we identified that the multi-channel CNN architecture manages to produce better validation accuracies compared to the accuracies produced by Recurrent Neural Network (RNN) based models, which are commonly used with sequence data.

4.1 Data

Emotion Classification We use the data from the 8th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA-2017). The data was used in the shared task to identify emotion intensity (Mohammad and Bravo-Marquez, 2017). The tweets in the dataset were assigned with the labels: anger, fear, joy and sadness, and their associated intensities. Table 1 presents the detailed statistics of the dataset.

Emotion	Train	Test	Dev	Total
Anger	857	760	84	1701
Fear	1147	995	110	2252
Joy	823	714	79	1616
Sadness	786	673	74	1533
Total	3613	3142	347	7102

Table 1: The number of tweets under each emotion category

The dataset contains 194 tweets that belong to multiple emotion categories. For example, the tweet: "I feel like I am drowning. #depression #anxiety #falure #worthless" is associated with the labels 'fear' and 'sadness'. When training the model, we created a single training dataset by combining both the train and test data and tested the trained model on the development dataset. The main reason for using such an approach is to im-

prove the neural network model training by providing as much data as possible so that model overfitting will be reduced while increasing the model generalization. During our training, we did not take into consideration the emotional intensity and expect to use it in our future research as an additional input.

Mental Illness Detection To detect whether a user is a neurotypical user or if the user is susceptible to having either PTSD or depression, we used the dataset from the Computational Linguistics and Clinical Psychology (CLPsych) 2015 shared task (Coppersmith et al., 2015b). Table 2 presents the detailed statistics of this dataset.

	Control	**PTSD**	**Depressed**
Number of users	572	246	327
Average age	24.4	27.9	21.7
Gender (female) distribution per class	74%	67%	80%

Table 2: CLPSych 2015 shared task dataset statistics

Preprocessing: All the URLs, @mentions, #hashtags, RTweets, emoticons, emojis, and numbers were removed. We removed a selected set of stopwords but kept first, second, and third person pronouns. It was discovered that users susceptible to mental disorders such as depression have frequently used the first-person singular pronouns compared to neurotypical users (Pennebaker et al., 2007). Also, the punctuation marks except for a selected few were removed. The full stop, comma, exclamation point, and the question mark were kept while removing all the other punctuation marks. The NLTK (Natural Language Toolkit) tweet tokenizer was used to tokenize the tweets. We selected 200,000 unique tokens to build the vocabulary, rather than choosing all the unique words, which could lead to sparse word vectors with high dimensionality.

Vocabulary Generation: To obtain an enriched dictionary containing the most relevant terms, we introduced a novel approach instead of the traditional approach used in many deep learning APIs (e.g., Keras deep learning high-level API[1]). Our approach takes into account the top 'K' terms based on their term frequency and inverse document frequency (TF-IDF) scores. To build the dictionary, first, we calculated the TF-IDF values under each user (i.e., by considering all the train/validation tweets of a single user as one single document). Then we took the maximum score out from all the assigned TF-IDF scores given the word. The reason for taking the maximum is to extract the words identified as closely related to a given user. Based on the computed TF-IDF scores, we picked the top 'K' words (K=200,000) to construct the vocabulary. A dictionary created using the above approach allows the model to capture the underlying relationships between the critical words. In comparison to the word frequency-based approach, using the vocabulary based on the TF-IDF scores has produced relatively better results for the recorded matrices (refer Table 4). When analyzing the model's prediction accuracy and loss (i.e., on training and validation data) over five-fold cross-validation, we identified that the model trained using the TF-IDF based vocabulary has been more stable with less randomness compared to the model trained using the vocabulary based on word frequencies.

When choosing the maximum sequence length for the input data, it is essential to capture as much information as possible from each user, especially given consideration to the research domain of mental illness detection. Since we have concatenated all the individual tweets belonging to one user as a single string, a high variance in the sequence length was identified among users. On average, a single user has used around 15,800 tokens, where the maximum number of tokens used by an individual user is nearly 64,800. Rather than experimenting with different sequence lengths, we selected the maximum length for the sequence by adding three standard deviations to the average sequence length covering 99% of users with a sequence length of 46,200 tokens. The shorter sequences were padded with zeros (to the end of the sequence), and the longer sequences were truncated (from the end of the sequence).

Model Architecture The selected model architecture consists of three main components: multi-task learning, CNN with multi-channel, and multi-inputs. Multi-task learning is known to be successful when the data is noisy and limited so that when trying to learn one task, it could gain additional information from the other tasks to identify the most relevant features. Learning a shared representation so that individual tasks can benefit

[1]https://keras.io/preprocessing/text/

from one another (Caruana, 1997) can be considered as one of the most appropriate architectures when trying to detect multiple mental illnesses. Benton et al. (2017b) demonstrated the successful use of multi-task learning to recognize mental illnesses and suicide ideation. Different from their approach, we add multiple features discovered by researchers in the fields of computational linguistics and psychology, to enhance the model performances. We consider that it is important to identify the impact of manually engineered features on the model's predictability. We also recognized that using a CNN multi-channel architecture is best suited for tasks dealing with limited unstructured data compared to RNN architectures or multilayer perceptrons (MLP).

We used a multi-channel model as the base model in both emotion classification (i.e., to detect anger, sadness, joy, fear) and mental illness detection (i.e., to detect PTSD and depression). The multi-channel model uses three versions of a standard CNN architecture with different kernel sizes. We identified that using different kernel sizes (different n-grams sizes) with Global Maximum Pooling produces better results compared to a standard CNN architecture. The optimal validation accuracies for both emotion and mental illness detection models were derived using three channels with kernel sizes 1, 2, and 3. Increasing the kernel sizes or the number of channels reduced the validation accuracies.

For the emotion classification task, the CNN in each channel was tested with 64 filters, same padding and a stride of 1 (distance between successive sliding windows). We used Rectified Linear Unit (ReLU) as the activation function. To normalize the data and to reduce the impact of model overfitting, we used the batch normalization layer and used a dropout (Srivastava et al., 2014) as the regularization technique with a probability of 0.2. As the final layer in each channel, we used global maximum pooling to reduce the number of parameters needed to learn so that it could further reduce the impact of model overfitting. The outputs from each global maximum pooling layers (from each channel) were concatenated and fed into a fully connected layer with four hidden units that use sigmoid activation to generate the output. All the inputs were sent through trainable embedding layers (randomly initialized) with a dimension of 300 for the emotion classification task and 100 for the mental illness detection task.

Throughout our research, we did not emphasize much on word embeddings as our primary objective was to identify the impact of merging features derived using deep learning methods with few of the notable features that were identified over the years by researchers on detecting mental illnesses. Even though our best results were obtained by instantiating the embedding layer weights with random numbers (refer Table 4), we conducted several preliminary experiments using word embeddings trained on the fastText (Joulin et al., 2017) algorithm. We decided to use fastText because given the unstructured nature of the twitter messages we could obtain a more meaningful representation by expressing a word as a vector constructed out of the sum of several vectors for different parts of the word (Bojanowski et al., 2017). One of the reasons for the low measurements could be due to the reason that we used fewer data to train our embeddings.

Mental illness detection When building the multi-task learning model to detect mental illnesses, the base model architecture (i.e., for the shared representation) has used a structure similar to the one used in emotion classification. The fundamental changes to the base model include: using 256 filters instead of 64 and using L1 kernel regularization in each convolution layer. We used the trained model on the emotion data to predict the emotion category of the individual Twitter messages in the CLPSych 2015 dataset. We grouped the predicted probabilities for each user under the different emotion categories by calculating the standard deviation. We used the predicted probabilities for each emotion category as multiple inputs when detecting neurotypical and depressed users, while age and gender were used as inputs when predicting users with PTSD. Before concatenating the multiple-inputs with the output from the multi-channel architecture, the multiple-inputs were transformed using a fully connected layer with 128 hidden units and ReLU activation. The output from the shared layers and the transformed multiple inputs were merged before being used as the input to the fully connected layer with three individual hidden units and sigmoid activation. Before applying multiple-inputs to the neural network architecture, all the relevant inputs were normalized using a minimum, maximum scaler initialized within the range 0 and 1. The neu-

ral network architecture used for the multi-task, multi-channel, multi-input model for mental illness detection is shown in Figure 1.

Model Training When training both the emotion and the mental illness detection models, we minimized the validation loss to learn the optimal neural network parameters. To train both the models, we used minibatch gradient descent with smaller batch sizes. Using a smaller batch size is known to stabilize model training while increasing the model generalizability. In many cases, the optimal results were obtained by using batch sizes smaller or equal to 32 (Masters and Luschi, 2018). In our experiments, we used batch sizes 32 and 8 respectively for training the emotion and mental illness detection models. Both models were trained for 15 epochs and used early stopping when the validation loss has stopped improving. The Adam optimizer (Kingma and Ba, 2014) was used when training both the models with the default learning rate of 0.001.

5 Results

5.1 Emotion classification

The emotion detection task was implemented as a multi-class, multi-label classification because the same Twitter message can belong to multiple classes. Since we would like to have independent probability values for each class rather than a probability distribution over the four classes, we used binary cross-entropy as the loss function. Having independent probability values is better-oriented towards the identification of independent emotion categories. Table 3 reports the emotion classification results obtained using multi-channel CNN (MCCNN), and in addition, results from several other experiments: CNN with max-pooling (CNNMax) and bidirectional Long Short Term Memory module (biLSTM) were reported for comparison.

	Acc (%)	F1(%)	P(%)	R(%)	P rank (%)
MCCNN	88.88	77.41	79.68	75.67	84.85
CNNMax	85.82	68.95	76.97	62.70	81.39
biLSTM	85.07	68.66	75.97	63.49	80.97

Table 3: Emotion multi-class, multi-label classification results

In Table 3, the recorded accuracy is based on the Keras API[2] accuracy calculation on the multi-class, multi-label models where it takes into account the individual label predictions rather than a complete match (i.e., if there is more than one label per instance). The F1-score, precision, and recall measures are calculated based on the 'macro' averaging on the exact match and hence the low percentages. We have also reported the label ranking average precision score, which averages over the individual ground truth label per instance. This metric is mainly used in tasks that involve multi-label ranking. From the reported results, it can be seen that the multi-channel CNN model has given the best results compared to the standard CNN model and the RNN based model. Based on the outcome, we have used the above trained multi-channel CNN model to make predictions on the CLPsych 2015 individual tweets.

5.2 Mental illness detection

For detecting mental illnesses, we used binary cross-entropy loss as the loss function and sigmoid activation as the final layer activation. The data was sampled using Stratified Shuffle Split to maintain class distribution between 80%/20% of training and validation data. Our models were evaluated only on the validation data because the CLPSych 2015 shared task test data labels were not made available. To ascertain the model reliability, we performed 5-fold cross-validation and discovered that having multi-inputs on a multi-task, multi-channel architecture does increase the model performances. The recorded model accuracies were averaged over five folds with a standard deviation of 0.01.

Table 4 demonstrates the model performances according to different combinations of the multi-task, multi-channel, and multi-input architectures. To demonstrate the effectiveness of the proposed approach, we conducted several experiments using variants of two deep learning architectures; Convolutional Neural Networks, and Recurrent Neural Networks and also to measure a baseline, we used the shallow learning approach; Support Vector Machine. The experiments: MtMcMi (Multi-task Multi-channel, Multi-input), MtMc (Multi-task, Multi-channel), MtMcMiFT(Multi-task, Multi-channel, Multi-input, using FastText word representations), MtMcMiFr (Multi-task, Multi-channel, Multi-input, using word Frequen-

[2]https://keras.io/metrics/

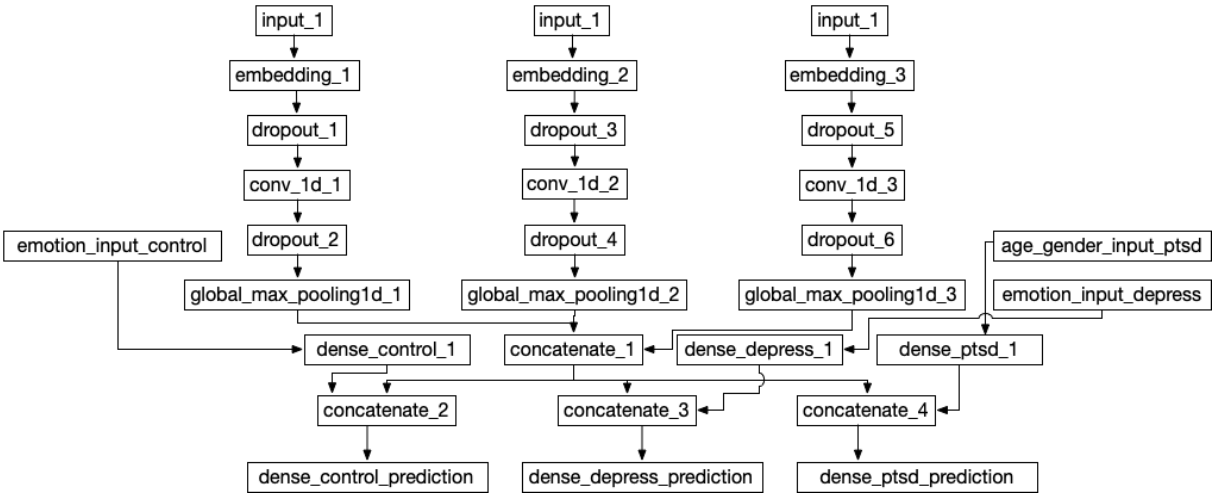

Figure 1: Multi-task, multi-channel, multi-input model for mental illness detection

	Accuracy(%)			**Avg. F1(%)**			**Avg. Precision(%)**			**Avg. Recall(%)**			**Avg. AUC(%)**		
	C	D	P	C	D	P	C	D	P	C	D	P	C	D	P
MtMcMi	89.08	**87.59**	91.35	**89.07**	**83.81**	86.06	**89.19**	**86.61**	89.81	89.07	**82.05**	83.50	**95.30**	**92.24**	**93.18**
MtMc	88.55	86.89	**91.96**	88.53	83.06	**87.02**	88.82	85.11	**90.54**	88.54	81.75	**84.64**	94.62	90.74	92.54
MtMcMiFT	85.50	86.28	91.26	85.45	82.73	85.90	85.91	84.06	89.70	85.49	81.97	83.30	93.88	91.01	91.91
MtMcMiFR	87.42	86.72	91.52	87.41	83.07	86.52	87.48	84.63	89.49	87.41	82.00	84.36	94.88	91.55	92.53
McMclass	**92.00**	75.69	76.73	89.05	76.83	81.22	86.34	78.25	86.93	**92.00**	75.69	76.73	92.44	84.55	86.32
biLSTMMtMi	51.35	71.61	78.60	41.27	41.73	44.00	41.92	35.80	39.30	51.52	50.00	50.00	56.87	59.89	60.28
biLSTMMt	52.48	72.31	78.60	47.48	46.40	44.00	47.84	59.81	39.30	52.57	52.06	50.00	56.32	59.96	54.89
svmMclass	81.73	52.91	42.85	75.82	57.01	46.87	70.76	62.11	51.97	81.73	52.91	42.85	81.18	79.12	77.70

Table 4: Mental illness detection using multi-task, multi-channel, multi-input architecture. The labels 'C', 'D' and 'P' denotes 'Control', 'Depressed' and 'PTSD'

cies), McMclass (Multi-channel, Multi-class), biLSTMMtMi (bidirectional Long Short Term Memory, Multi-task, Multi-input), biLSTMMt (bidirectional Long Short Term Memory, Multi-task), svmMclass (support vector machine, Multi-class) were conducted to identify a fitting approach to discover individuals who are susceptible of mental disorders.

The metrics used for the evaluation are accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic curve (AUC). The AUC score is used to compare the model performances where the average AUC is calculated with standard deviation. The standard deviation is used as a mechanism to identify variance among model performances, which could bring insight into the reliability of the trained model. For each experiment, we recognized that the standard deviation is approximately around 0.01, which provides an empirical confirmation that the sampling using stratified shuffle splits provides an accurate representation of the complete dataset.

The "MtMcMi" architecture uses features based on emotion as multi-inputs on the control and depressed users while age and gender were used on the users with PTSD. In comparison to "MtMc" which is multi-task, multi-channel without multi-inputs, we could see that using multiple inputs have increased the average AUC score, as well as most of the other evaluation matrices (i.e., precision, recall, and F1-score). Even though the increase could not be considered as significant, the potential room for improvement is high if provided with more accurate emotion prediction and additional profound features identified by researchers. Concerning the emotion detection task, we could have obtained considerable improvements in prediction accuracies if provided with additional data when training the deep learning model.

When analyzing the result for "MtMcMiFr", which is using the Multi-task, Multi-channel, Multi-input architecture with the vocabulary created using the default word frequencies; we could identify that our proposed approach, which uses the vocabulary constructed using the weighted TF-IDF words has produced comparatively better results. Even though the gained improvements could

not be considered as significant, the proposed method can be regarded as an effective approach when initiating the vocabulary, that provides a balance between the rare and frequently used words.

In the experiment "MtMcMiFT", where we used the fastText embeddings layer with the number of dimensions equal to 100 as an input to the Multi-channel convolutional neural network, it was observed that the results obtained using the randomly initialized embedding layer are higher than with the fastText pre-trained embeddings. This could be due to embeddings not being trained on a sufficiently large dataset (we trained them only on the CLPSych 2015 dataset). In future work, we will conduct further research to enhance the embedding layer word representation by using state-of-the-art language modeling approaches trained on larger datasets.

The effectiveness of using convolutional neural network models can be identified when evaluating the results obtained using Recurrent Neural Network (RNN) based architectures. The "biLSTMMtMi" method uses a bidirectional Long Short Term Memory ("biLSTM") model in a Multi-task, Multi-input design and comparatively has produced poor results for different combinations of hyperparameters. This could be due to several reasons such as the unstructured nature of the Twitter text as well as the non-existence of long-term dependencies. For example, our best results were obtained when using the kernel sizes (i.e., the number of consecutive tokens) one, two, and three and ones the kernel sizes are increased the overall model predictability decreases. When using a "biLSTM" model as the shared layer in multi-task learning without multiple inputs ("biLSTMMt"), the results are somewhat better compared to when using multi-inputs.

To demonstrate the effectiveness of using multi-task learning to detect multiple mental disorders, we compared the proposed approach with multiclass classification to distinguish neurotypical users from users susceptible to having either PTSD or depression. First, we used a multi-channel convolutional neural network to predict the three classes (i.e., control, depress, and PTSD). In comparison to our proposed approach, we can identify that multiclass classification using CNN have produced slightly better results on two occasions, which is for average accuracy and recall under the control class. Through further analysis, we see that average precision, F1-score, and AUC scores are higher for all three classes when using the proposed approach. Overall the multiclass classification task has produced low scores (especially for precision, F1-score, and AUC) when detecting users susceptible to depression and PTSD while the proposed approach has contributed significantly better results. The better results could be due to the reason that depression is commonly identified among individuals with PTSD, and the shared layer has managed to learn such common characteristics while the task-specific layers have learned the individual features unique to each disorder.

As a baseline, we used the linear SVM classifier with TF-IDF features (200,000 features) in a multiclass classification task (i.e., svmMclass). When sampling the data, five splits of 80% training and 20% testing were created using the Stratified Shuffle Split method to maintain class distribution. We also computed a majority class baseline, which classifies everything in the largest class (the control class in the training data). It achieved an accuracy of 50.21% on the test data. Overall, we can see that using limited unstructured data with an architecture based on CNN have produced better results compared to the solution based on RNN. Notably, the multi-task, multi-channel architecture with multiple-inputs has provided the best results and confirms that using multiple-inputs has a positive influence on the overall model performances. Also, the appropriateness of using multi-task learning instead of multiclass classification to detect multiple mental disorders is highlighted. Similar to the fact that certain mental disorders share specific common symptoms (American Psychiatric Association, 2013), multi-task learning has managed to learn such characteristics through a shared representation followed with task-specific layers to identify the unique attributes to differentiate multiple mental disorders.

6 Comparison to related work

Even though our work could not be directly compared with (Benton et al., 2017b), we can identify that our model has produced competitive results, especially when comparing the AUC score for detecting users with PTSD and depression. Our best model has scored an AUC score >0.90 in identifying all three individual classes (control, depression, PTSD) in comparison to an AUC

score <0.80 for detecting PTSD and depression and an AUC score >0.90 for detecting the neurotypical users (Benton et al., 2017b). In the CLPSych 2015 shared task (Coppersmith et al., 2015b), Resnik et al. (2015a) have reported AUC scores of 0.86 (depression vs. control), 0.84 (depression vs. PTSD) and 0.89 (PTSD vs. control) and similarly Preotiuc-Pietro et al. (2015) have reported an average AUC score around 0.86 in differentiating neurotypical users from users susceptible to PTSD and depression. Even though we have produced better results using the validation dataset, we could not directly compare our results with the shared task participants as they have evaluated their models against the test dataset which was not made available to us. In our future work, we will conduct experiments using public forum post data extracted from platforms such as Reddit[3]. In comparison, the proposed approach can be tested with adequate adjustments to detect multiple mental disorders such as depression, anxiety, PTSD, and six others using the dataset introduced by Cohan et al. (2018). The authors have achieved an F1-score of 27.83% for multi-class classification and 53.56% and 57.60% respectively, when detecting depression and PTSD as binary classification tasks.

7 Conclusion

In this paper, we investigated the impact of merging features derived using deep neural network architectures with profound manually engineered features identified by researchers over the years using shallow learning to detect mental disorders using social media text. In particular, we have identified that by using a multi-channel convolutional neural network as a shared layer in a multi-task learning architecture with multiple-inputs (e.g., different emotion categories, age, and gender) have produced comparatively competitive results in detecting multiple mental disorders (in our case depression and PTSD). For future work, we will continue our research on suicide risk detection, and the temporal impact different mental disorders have on suicide ideation.

References

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*. Arlington.

Sairam Balani and Munmun De Choudhury. 2015. Detecting and Characterizing Mental Health Related Self-Disclosure in Social Media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '15*, pages 1373–1378.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017a. Ethical Research Protocols for Social Media Health Research. In *First Workshop on Ethics in Natural Language Processing,*, pages 94–102.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017b. Multi-Task Learning for Mental Health using Social Media Text. *CoRR*, abs/1712.0.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135—-146.

Canadian Mental Health Association. 2016. Canadian Mental Health Association.

Rich Caruana. 1997. Multitask Learning. *Machine Learning*, pages 41–75.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485—-1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014a. Measuring Post Traumatic Stress Disorder in Twitter. In *In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM).*, volume 2, pages 23–45.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014b. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015a. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. In *Computational Linguistics and Clinical Psychology*, pages 1–10.

Glen Coppersmith, Mark Dredze, Craig Harman, Hollingshead Kristy, and Margaret Mitchell. 2015b. 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.

[3]https://www.reddit.com/

Munmun De Choudhury. 2013. Role of Social Media in Tackling Challenges in Mental Health. In *Proceedings of the 2nd International Workshop on Socially-Aware Multimedia (SAM'13)*, pages 49–52.

Munmun De Choudhury. 2014. Can social media help us reason about mental health? In *23rd International Conference on World Wide Web*, Cdc, pages 1243–1244.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Major Life Changes and Behavioral Markers in Social Media : Case of Childbirth. In *Computer Supported Cooperative Work (CSCW)*, pages 1431–1442.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013b. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, page 3267.

Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. *Cscw*, pages 626–638.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep Learning for Depression Detection of Twitter Users. In *Fifth Workshop on Computational Linguistics and Clinical Psychology*, pages 88–97.

Susan Jamison-Powell, Conor Linehan, Laura Daley, Andrew Garbett, and Shaun Lawson. 2012. "I can't get no sleep": discussing #insomnia on Twitter. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 1501–1510.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 427—-431.

Diederik P. Kingma and Jimmy Lei Ba. 2014. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pages 1–15.

Rohan Kshirsagar, Robert Morris, and Samuel Bowman. 2017. Detecting and Explaining Crisis. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 66–73, Vancouver. Association for Computational Linguistics.

Michael Thaul Lehrman, Cecilia Ovesdotter Alm, and Rubén A. Proaño. 2012. Detecting Distressed and Non-distressed Affect States in Short Forum Texts. In *Second Workshop on Language in Social Media*, Lsm, pages 9–18, Montreal.

Kate Loveys, Patrick Crutchley, Emily Wyatt, and Glen Coppersmith. 2017. Small but Mighty: Affective Micropatterns for Quantifying Mental Health from Social Media Language. *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology*, pages 85–95.

Dominic Masters and Carlo Luschi. 2018. Revisiting Small Batch Training for Deep Neural Networks. *CoRR*, pages 1–18.

Mental Health Commission of Canada. 2016. Mental Health Commission of Canada.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 Shared Task on Emotion Intensity. In *8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Minsu Park, David W Mcdonald, and Meeyoung Cha. 2013. Perception Differences between the Depressed and Non-depressed Users in Twitter. *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 476–485.

Umashanthi Pavalanathan and Munmun De Choudhury. 2015. Identity Management and Mental Health Discourse in Social Media Identity in Online Communities. In *WWW'15 Companion: 24th International World Wide Web Conference*, pages 18–22.

James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. 2007. The Development and Psychometric Properties of LIWC2007 The University of Texas at Austin. Technical Report 2, The University of Texas at Austin, Austin, Texas.

Daniel Preot, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. 2015. The Role of Personality , Age and Gender in Tweeting about Mental Illnesses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30.

Daniel Preotiuc-Pietro, Maarten Sap, H. Andrew Schwartz, and Lyle Ungar. 2015. Mental Illness Detection at the World Well-Being Project for the CLPsych 2015 Shared Task. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 40–45.

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 2015a. The University of Maryland CLPsych 2015 Shared Task System. In *CLPsych 2015 Shared Task System*, c, pages 54–60.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-an Nguyen, and Jordan Boyd-graber. 2015b. Beyond LDA : Exploring Supervised Topic Modeling for Depression-Related

Language in Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, volume 1, pages 99–107.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using Topic Modeling to Improve Prediction of Neuroticism and Depression in College Students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, October, pages 1348–1353.

H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards Assessing Changes in Degree of Depression through Facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Statista. 2017. Number of social media users worldwide from 2010 to 2021.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing Depression from Twitter Activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 3187–3196.

World Health Organization. 2014. WHO — Mental health: a state of well-being.

World Health Organization. 2018a. Depression.

World Health Organization. 2018b. Mental disorders.