Deep Learning for Depression Detection of Twitter Users

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, Diana Inkpen
School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, ON K1N 6N5 Canada
{ahuss045, pkiri056, mhuss092, diana.inkpen}@uottawa.ca

Abstract

Mental illness detection in social media can be considered a complex task, mainly due to the complicated nature of mental disorders. In recent years, this research area has started to evolve with the continuous increase in popularity of social media platforms that became an integral part of people's life. This close relationship between social media platforms and their users has made these platforms to reflect the users' personal life on many levels. In such an environment, researchers are presented with a wealth of information regarding one's life. In addition to the level of complexity in identifying mental illnesses through social media platforms, adopting supervised machine learning approaches such as deep neural networks have not been widely accepted due to the difficulties in obtaining sufficient amounts of annotated training data. Due to these reasons, we try to identify the most effective deep neural network architecture among a few of selected architectures that were successfully used in natural language processing tasks. The chosen architectures are used to detect users with signs of mental illnesses (depression in our case) given limited unstructured text data extracted from the Twitter social media platform.

1 Introduction

Mental disorder is defined as 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” (American Psychiatric Association, 2013). According to Canadian Mental Health Association (2016), 20% of Canadians belonging to different demographics have experienced mental illnesses during their lifetime, and around 8% of adults have gone through a major depression. According to World Health Organization (2014) statistics, nearly 20% of children and adolescents have experienced mental illnesses and half of these mental illnesses start before the age of 14. In addition, around 23% of deaths in the world were caused due to mental and substance use disorders. The broad implication of mental illness can be identified from the level of suicide in Canada where nearly 4,000 Canadians have died from suicide and 90% of them were identified as having some form of a mental disorder (Mental Health Commission of Canada, 2016). Apart from the severity of mental disorders and their influence on one's mental and physical health, the social stigma (e.g., ”mental disorders cannot be cured”) or discrimination has made the individuals to be neglected by the community as well as to avoid taking the necessary treatments.

The inherent complexity of detecting mental disorders using social media platforms can be seen in the literature, where many researchers have tried to identify key indicators utilizing different natural language processing approaches. To extract the most prominent features to develop an accurate predictive model, one must acquire a sufficient amount of knowledge related to the particular area of research. Even if such features were extracted, this does not assure that those features are the key contributors to obtaining improved accuracies. Due to these reasons, we investigate the possibility of using deep neural architectures because the features are learned within the architecture itself.

Here, we explore a few selected deep neural network architectures to detect mental disorders, specifically depression. We used the data released for the Computational Linguistics and Clinical Psychology (CLPsych) 2015 shared task (Coppersmith et al., 2015b). Even though the task is comprised of three subtasks: detecting Post-Traumatic

Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 88–97
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

Stress Disorder (PTSD) vs. control, depression vs. control and PTSD vs. depression, our primary objective was to detect depression using the most effective deep neural architecture from two of the most popular deep learning approaches in the field of natural language processing: Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), given the limited amount (i.e., in comparison to most of the deep neural network architectures) of unstructured data.

Our approach and key contributions can be summarized as follows.

- Word embedding optimization: we propose a novel approach to optimize word-embedding for classification with a focus on identifying users suffering from depression based on their social posts such as tweets. We use our approach to improve the performance of two tasks: depression detection on the CLPsych2015 dataset and test generalization capability on the Bell Lets Talk dataset (Jamil et al., 2017).
- Comparative evaluation: we investigate and report the performance of several deep learning architectures commonly used in NLP tasks, in particular, to detect mental disorders. We also expand our investigation to include different word embeddings and hyperparameter tuning.

2 Mental illness detection

Identifying the treatment requirement for a mental disorder is a complicated clinical decision, which involves several factors such as the severity of symptoms, patients' suffering associated with the symptoms, positive and negative outcomes of particular treatments, disabilities related to patients' symptoms, and symptoms that could negatively impact other illnesses (American Psychiatric Association, 2013). It is also important to note that measuring the severity of the disorder is also a difficult task that could only be done by a highly trained professional with the use of different techniques such as text descriptions and clinical interviews, as well as their judgments (American Psychiatric Association, 2013). Considering the complexity of the procedures and level of skills involved in identifying mental disorder and the necessary treatments, detecting mental illness within social media using web mining and emotion analysis techniques could be considered a preliminary step that could be used to generate awareness.

It is of greater concern to respect ethical facets about the use of social media data and its privacy. The researchers working with such social media data must take the necessary precautions to protect the privacy of users and their ethical rights to avoid further psychological distress. Certain researchers have taken adequate steps in anonymizing the data to secure user privacy. Coppersmith et al. (2015b) have used a whitelist approach in anonymizing the data given to the CLPsych 2015 shared task participants. Even though screen names and URLs were anonymized using salted hash functions, the possibility of cross-referencing the hashed text against the Twitter archives still exists, and it could lead to breach of user privacy. Due to this reason, the researchers were asked to sign a confidentiality agreement to ensure the privacy of the data.

As social media interactions reside in a more naturalistic setting, it is important to identify to what extent an individual has disclosed their personal information, and whether the accurate and sufficient information is being published to determine whether a person has a mental disorder. The longitudinal data published on social media platforms have been identified as valuable (De Choudhury, 2013, 2014, 2015) with an extensive level of self-disclosure (Balani and De Choudhury, 2015; Park et al., 2012).

Most of the research conducted to detect mental illnesses in social media platforms has focused heavily on feature engineering. Throughout the literature, it could be identified that the most widely adopted feature engineering method is to extract lexical features using the Linguistic inquiry word count (LIWC) lexicon, which contains more than 32 categories of psychological constructs (Pennebaker et al., 2007). The lexicons have been used as one of the key feature extraction mechanisms in identifying insomnia (Jamison-Powell et al., 2012), distress (Lehrman et al., 2012), postpartum depression (De Choudhury et al., 2013), depression (Schwartz et al., 2014) and post-traumatic stress disorder (PTSD) (Coppersmith et al., 2014a). For each of these mental disorders to be identified, researchers had to extract features that overlap with each other, and are unique to a particular disorder. For example, the use of first-person pronouns (Lehrman et al., 2012) compared to the lesser use of second and third person pronouns (De Choudhury, 2013) are being used to detect users susceptible to distress and depression. To distinguish depression from PTSD, age is identified as a distinct feature (Preotiuc-Pietro et al., 2015a).

We found that working with the data extracted from the Twitter social media platform is challenging due to the unstructured nature of the text posted by users. The Twitter posts are introduced with new terms, misspelled words, syntactic errors, and character limitations when composing a message. Character n-gram models could be considered as an intuitive approach to overcome challenges imposed by unstructured data. Considering the effectiveness of such language models in classification tasks using Twitter data, Coppersmith et al. (2014a,b) has used unigram and character n-gram language models to extract features in the process of identifying users suspicious of having PTSD and several other mental illnesses such as bipolar disorder, depression, and seasonal affective disorder (SAD). Similarly, character n-grams can be identified as the key feature extraction mechanism in detecting mental illnesses such as attention deficit hyperactivity disorder (ADHD), generalized anxiety disorder, and eight other mental illnesses (Coppersmith et al., 2015a) as well as in detecting rare mental health conditions such as schizophrenia (Mitchell et al., 2015). Even though topic modelling techniques such as latent Dirichlet allocation (LDA) are being used to enhance the classifier predictability (Mitchell et al., 2015), researchers have identified supervised topic modeling methods (Resnik et al., 2015) and topics derived from clustering methods such as Word2Vec and GloVe Word Clusters (Preotiuc-Pietro et al., 2015b) to be more reliable in identifying users susceptible to having a mental illness. Further advancements in detecting mental health conditions were identified in the Computational Linguistics and Clinical Psychology (CLPsych) 2016 shared task (Milne et al., 2016) where post embedding's (Kim et al., 2016) were used to determine the category of severity (i.e., crisis, red, amber and green) of forum posts published by users. In addition to lexical (e.g., character n-grams, word n-grams, lemma n-grams) and syntactic features (e.g., POS n-grams, dependencies), social behavioural patterns such as posting frequency and retweet rate, as well as the demographic details such as age, gender, and personality (Preotiuc-Pietro et al., 2015a) were also considered strong indicators in identifying mental illnesses. In general, the research in mental illness detection has evolved from the use of lexicon-based approaches to language models and topic models. The most recent research has tried to enhance models' performance with the use of vector space representations and recurrent neural network layers with attention (Kshirsagar et al., 2017) to detect and explain posts depicting crisis. In our research, we implement a model that produces competitive results for detecting depression of Twitter users (i.e., at user level not at post level) with limited data and without any exhaustive feature engineering.

	Control	**Depressed**	**PTSD**
Number of users	572	327	246
Number of tweets in each category	1,250,606	742,793	544,815
Average age	24.4	21.7	27.9
Gender (female) distribution per class	74%	80%	67%

Table 1: CLPSych 2015 shared task dataset statistics

3 Data

The training data consists of 1,145 Twitter users labeled as Control, Depressed, and PTSD (Coppersmith et al., 2015b). Also, each user of the dataset is labeled according to their gender and age. Table 1 represents detailed statistics of the dataset.

As the research is focused mainly on identifying users susceptible to depression, we selected a test dataset consisting 154 users labeled as either Depressed or Control. The users are identified from the postings published under the Bell Let's Talk campaign (Jamil et al., 2017). Out from 154 users, 53 users are labeled as Depressed while the remaining 101 users as Control. The test dataset can be considered as random and not following the same distribution as the training dataset. The training data contained an average length of 13,041 words per user, and on average 3,864 words are used by a user in the test set. Unlike the training dataset, the test dataset is not extracted considering the age and gender attributes, and it does not have a similar age and gender distribution between the control and depressed groups. We assume that our trained model could generate better AUC scores if provided with a similarly distributed test dataset. However, considering the

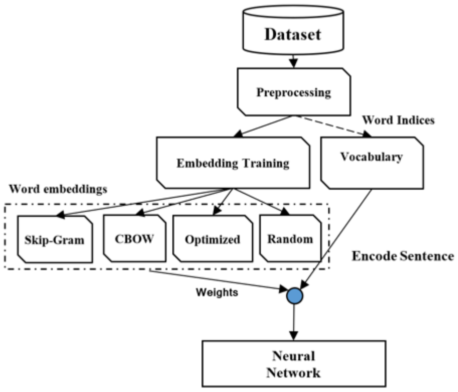

Figure 1: System architecture.

AUC scores that we have obtained, we can conclude that the trained model is well generalized.

4 Methodology

The overall design of our approach is shown in Figure 1. We present a system that identifies users at the risk of depression from their social media posts.

Toward this, we present an efficient neural network architecture that improves and optimizes word embeddings. We evaluate the optimized embeddings produced by our architecture along with three commonly used word embeddings (Figure 1), random trainable, skip-gram, and CBOW, on the CLPsych 2015 shared task and the Bell Let's Talk datasets. We perform a comparison on some selected CNN-based and RNN-based models to determine the best models and parameters across different settings for depression detection.

4.1 Preprocessing

We removed all the retweets, URL's, @mentions, and all the non-alphanumeric characters. Also, all the stop words except for first, second, and third person pronouns were removed. From previous research, we identified that individuals susceptible to depression more regularly use first-person singular pronouns compared to the use of other pronouns (Pennebaker, 2011). The NLTK Tweet tokenizer is used to tokenize the messages. After tokenizing, we build a vocabulary (242,657 unique tokens) from the training dataset, which is used to encode text as a sequence of indices.

4.2 Word encoding

A network input is a sequence of tokens, such as words, where $S = [s_1, s_2, \ldots, s_t]$ and t denotes the timestep. S_i is the one-hot encoding of input tokens that have a fixed length (T), such that a sequence that exceeds this length is truncated. A word dictionary of fixed terms W is used to encode a sequence. It contains three constants that determine the start and end of this sequence, in addition to the out of vocabulary (OOV) words. We normalize the variable text length using padding for short sequences and truncation for long sequences. We set the minimum occurrences of a word to 2 and the size of context window to 5, which produce 242,657 words. Then, we select the most frequent 100,000 words of them without stopwords.

4.3 Word Embedding Models

Word embedding models are fundamentally based on the unsupervised training of distributed representations, which can be used to solve supervised tasks. They are used to project words into a low-dimensional vector representation x_i, where $x_i \epsilon R^W$ and W is the word weight embedding matrix. We pre-train two different Word2Vec (Mikolov et al., 2013) word embeddings, using Skip-gram and Continuous Bag-Of-Words (CBOW) distributed representation, in addition to a random (Rand) word embedding that has a uniform distribution scheme of a range (-0.5 to +0.5).

Word2Vec is a shallow model, in which neural layers, typically two, are trained to reconstruct a word context or the current word from their surrounding window of words. Skip-gram infers the nearby contextual words, as opposed to other distributed representations, such as CBOW, that focus on predicting current words. CBOW is a continuous skip-gram, in which the order of context words does not affect prediction or projection. CBOW is typically faster than skip-gram, which is slower but able to identify rare words (Mikolov et al., 2013).

Our embedding models are pre-trained on the CLPsych 2015 Shared task data. We also have an additional hyperparameter that is used to either freeze the embedding weight matrix or allow for further training.

4.4 Word Embedding Optimization

We implement an optimized approach for building an efficient word embedding to learn a better feature representation of health-specific tasks. Recently, there has been an increased use of *embeddings average* to compute word embedding, which

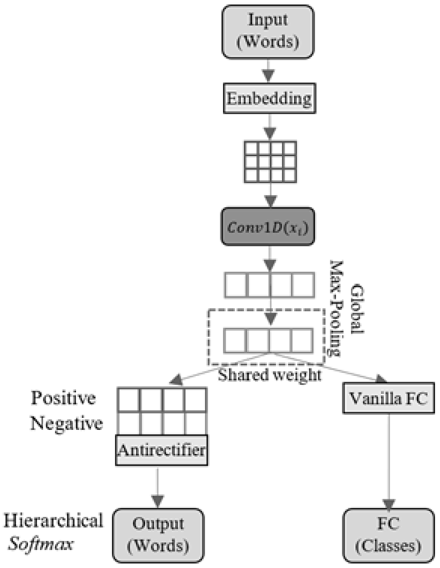

Figure 2: Word embedding optimization.

provides an improved feature representation that can be used across multiple tasks (Faruqui et al., 2015).

A word embedding is typically trained in an unsupervised manner using unlabeled data since it is not task-specific. We do the same by training our embedding on a large unlabeled training corpus (Collobert and Weston, 2008; Mikolov et al., 2013). Word2Vec trains word embeddings in a supervised manner, as it defines a training criterion that enables using unlabeled data (e.g., predicting the current word as in CBOW or context as in skip-gram). We do the same at the sentence level by predicting the surrounding sentences (Hill et al., 2016), as well as their possible sense (i.e., depressed, PTSD, or neither). We extend it by leveraging our knowledge about the labels of some sentences, where it will improve the estimation of word embedding and produces embeddings that are general-purpose and can be used across multiple tasks. We use multi-task deep learning (MTL) (Collobert and Weston, 2008) to learn word embedding by exploiting our knowledge of some labeled text as illustrated in Figure 2 (the shared layer among these tasks is in a dashed box).

Training. We have two tasks to be trained, word and sense predictions. We use a pre-trained weight matrix, in particular, skip-gram, to initialize the input word embedding. For the first task, we use supervised training to predict words occurring together (i.e., a pair of words w_i and w_j). For the second task, there is a fully-connected layer with Rectified Liner Unit (ReLU) activation, and a final layer producing the output. It includes a label for missing data, as it is expected to have limited supervised data regarding sense information. Then, we use a regularized l_2-norm loss function (Ng, 2004) to constrain shared layers between these tasks (see the dashed box in Figure 2).

For the first task, we define a probability $p(w_i, w_j) = e^{(cos(w_i,w_j))}/\Sigma_{w_i \epsilon W} e^{cos(w_i,w_j)}$ for the likelihood of word to be adjacent using a hierarchal SoftMax function. w_i denotes an embedding of a word w_i. W is the set of all possible words, many of which may not be practical

Hence, we replace the set W with the union of the sets W^C, W^P and W^N.W^C denotes the class index; depressed, PTSD, neither, or unknown. W^P denotes the words occurring next to a word w_i in the training data. W^N is a set of n words that are randomly selected and not occurring next to the word w_i in the training data. We use an antirectifier activation as it enables all-positive outputs without losing any value. Then, we use cosine distance function to compute similarities among word representations, and to produce word probability representations.

5 Models

We describe four selected neural network models, which are used to evaluate the performance of depression detection. The first three models use CNN and the last one uses RNN. We build these model on the top of the word-embeddings described in the previous section. A drop-out of a probability 0.2 follows the word embedding layer. Each model is followed by a vanilla layer that is fully-connected, has 250 hidden units, and uses a Rectified Linear Unit (ReLU) activation. Then, we apply dropout with a probability of 0.2. The output layer is a fully-connected layer with one hidden unit, and it uses a sigmoid activation to produce an output.

5.1 Convolutional Neural Network (CNN)

A convolution operation is a representation of learning from sliding w-grams for an input sequence of d entries, $e_1, e_2, \dots, e_t$. A vector $c_i \epsilon R^{ed}$ is the concatenated embedding of f entries, such that $x_{i-f+1}, \dots, x_i$ where f is the filter length. For w-gram, we generate a representation $p_i \epsilon R^d$ using convolution weights $W \epsilon R^{d \times wd}$ where a bias $b \epsilon R^d$ and $p_i = tanh(W_{x_i+b})$.

CNNWithMax: We apply a one-dimensional convolution operation with 250 filters and a kernel of size 3, where $w_i^f = conv1d(s_i)$ and f is the filter length. After that, a global max-pooling layer is applied on the feature map to extract global abstract information, such that $\widehat{w^f} = globalmax(w_i^f)$, which results in an abstract feature representation of length 250.

MultiChannelCNN: We apply 3 convolutions, each of which has 128 features and filters of the lengths 3, 4, and 5. A one-dimensional operation is used, where $w_i^f = Conv1d(S_i)$, and f is the filter length. Then, a max-pooling layer is applied on the feature map to extract abstract information, $\widehat{w_i^f} = max(C_i^f)$. Finally, we concatenate feature representations into a single output. Conversely to recurrent layers, convolutional operations are helpful with max-pooling to extract word features without considering the sequence order (Kalchbrenner et al., 2014). Such features can be used with recurrent features in order to improve the model performance.

MultiChannelPoolingCNN: We extend the previous model to apply two different max-pooling sizes, 2 and 5.

5.2 Recurrent Neural Network (RNN)

It is commonly used in NLP as it allows for remembering values over different time durations. In RNN, each element of an input embedding x_i is processed sequentially. $h_t = tanh(W_{x_i} + W_{h_{t-1}})$ and W represent the weight matrix between an input and hidden states (h_t) of the recurrent connection at timestep (t). RNN allows for variable length processing while maintaining the sequence order. However, it is limited when it comes to long sentences due to the exponentially growing or decaying gradients. Long short term memory (LSTM) is a common way to handle such a limitation using gating mechanisms.

Bidirectional LSTM with attention: we use bidirectional LSTM layers with 100 units, which receive a sequence of tokens as inputs. Then, the LSTM projects word information $H = (h_1, h_2, \ldots, h_T)$, in which h_t denotes the hidden state of LSTM at a timestep (t). LSTM captures the temporal and abstract information of sequences forwardly (h^f) or backwardly (h^b). Then, we concatenate both forward and backward representations, where $h_t = {h_t}^f || {h_t}^b$. Finally, we use the last output in the sequence.

Context-aware Attention: Words have different weight values, as they are generally not equal. Thus, we use an attention mechanism (Bahdanau et al., 2014; Vaswani et al., 2017) to focus on the important words. We use a context-aware attention mechanism (Yang et al., 2016), which is the weighted summation of all words in a given sequence $(r = \Sigma_{i=1}^{T} a_i h_i)$. We use this representation as a classification feature vector.

6 Models Training

For training, we minimize the validation loss error between the actual and predicted classes in order to learn the network parameters. A mini-batch gradient descent with a batch size 32 is applied to improve the network loss function through backpropagation. Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.007 is used to train our models. The gradient norm (Pascanu, Mikolov, & Bengio, 2012) is clipped at 7, which protects our model from the exploding gradient.

Regularization: we randomly drop neurons off a network using dropout in order to prevent co-adaptation of those neurons (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). Dropout is also used on the recurrent connection of our LSTM layers. We additionally, apply weight decay using L2 regularization penalty (Cortes, Mohri, & Rostamizadeh, 2012).

Hyperparameters: we use an embedding layer of the size 300, and an LSTM layer of size 50, which increases to be 100 for the bidirectional LSTM. We apply a dropout of 0.4 and 0.2 on the recurrent connections. Finally, an L2 regularization of 0.0001 is applied at the loss function.

7 Experiments

We evaluate our approach using two experiments, 1) depression detection on the CLPsych2015 dataset (Section 3) and 2) test generalization ability on the Bell Let's Talk dataset (Section 3). We use different word embeddings for our experiments with the deep neural network models. In the first experiment, we perform a comparison on the selected models for depression detection.

In the second experiment, since the dataset is imbalanced, we perform 5-fold cross-validation with stratified sampling to report results. Data points are shuffled for each split while maintaining the class distribution. After that, we test the generalization ability of the models selected, for which we use 80% and 20% of the data for train-

Model		Accuracy	F1	AUC	Precision	Recall
Baseline		77.480	77.472	0.844	77.601	77.480
CNNWithMax	**Optimized**	**87.957**	**86.967**	**0.951**	**87.435**	**87.029**
	Skip-gram	79.813	78.460	0.879	79.707	78.979
	CBOW	60.768	43.095	0.544	38.056	54.207
	Trainable	80.820	80.173	0.909	80.440	82.099
MultiChannelPoolingCNN	**Optimized**	87.510	86.491	0.950	87.266	86.678
	Skip-gram	78.818	76.073	0.883	80.514	75.691
	CBOW	49.667	37.573	0.556	33.289	53.652
	Trainable	73.691	72.021	0.824	72.672	72.799
MultiChannelCNN	**Optimized**	85.617	84.153	0.935	85.817	84.064
	Skip-gram	81.161	78.650	0.892	81.143	77.977
	CBOW	76.248	72.047	0.803	76.478	71.742
	Trainable	82.268	80.347	0.870	82.770	79.983
BiLSTM (Context-aware attention)	**Optimized**	78.136	76.024	0.826	76.555	75.751
	Trainable	77.589	75.193	0.832	76.687	74.923

Table 2: Performance of our models on the CLPsych 2015 dataset with 5-fold cross-validation. The rows are highlighted according to the highest AUC score.

Model		Accuracy	F1	AUC	Precision	Recall
Baseline		73.460	73.460	0.718	73.322	74.025
CNNWithMax	**Trainable**	64.935	64.787	0.751	68.376	69.681
	Optimized	81.818	80.998	0.920	80.529	83.449
	CBOW	61.688	61.216	0.687	63.214	64.515
	Skip-gram	72.078	71.322	0.743	71.879	74.229
MultiChannelCNN	**Trainable**	68.182	67.456	0.773	68.387	70.362
	Optimized	**83.117**	**82.252**	**0.923**	**81.626**	**84.439**
	CBOW	72.078	66.882	0.734	68.969	66.159
	Skip-gram	62.338	57.491	0.586	57.687	57.388
MultiChannelPoolingCNN	**Trainable**	60.390	54.599	0.525	54.911	54.558
	Optimized	82.468	81.513	0.888	80.871	83.495
	CBOW	51.948	50.752	0.682	69.076	62.918
	Skip-gram	64.286	64.248	0.752	69.307	70.082
BiLSTM (Context-aware attention)	**Trainable**	63.636	62.731	0.733	63.636	65.104
	Optimized	80.519	80.035	0.914	80.519	83.803

Table 3: Performance of our models on the Bell Let's Talk dataset. The rows are highlighted according to the highest AUC score.

ing and development, respectively. The trained models are used afterward for evaluation on unseen data, which is Bell Let's Talk; i.e., 154 users (Section 3).

The metrics used for our evaluation are accuracy, ROC area-under-the-curve (AUC), precision, recall, and F-measure. We use precision and recall since data is imbalanced, which may return imprecise accuracy results. We compared model performances based on the AUC score, which is calculated on the validation set and averaged over the five splits with standard deviation. A low precision will be identified when the classifier reports more false positives (FP); i.e., users are inaccurately predicted to have depression. A low recall will be identified when the classifier reports more false negatives (FN); i.e., users who suffer from depression are not recognized. We consider precision, recall, and F-measure for the positive classes obtained from the test datasets. We aim to be close to a perfect balance (1.0) for both precision and recall.

The majority of the researchers have relied on support vector machine (SVM) classifiers to distinguish users with mental disorders from control groups and different mental disorder categories except when trying to identify the level of depression with the use of regression models (Schwartz et al., 2014). We used the SVM linear classifier with TF-IDF to initiate a baseline for the binary classification task. For evaluation, we used five-fold cross-validation, and the resulting best model was used on the Bell Let's Talk dataset to predict users with depression. The results are reported both on the validation and test data.

Table 2 shows good standings results for depression detection, which indicates that regularization and hyperparameter tuning helped resolve the overfitting issues. CNN-based with max-pooling models reported better performance than RNN-based models. The CNNWithMax models using our optimized embedding reported higher accuracy (87.957%), F1 (86.967%), AUC (0.951), precision (87.435%), and recall (87.029%), as compared to other models. Table 2 reports that CNN-based models' results are close to each other, as opposed to RNN-based models, which at best reported 83.236% with trainable random embedding (trainable). Interestingly, CNN models performed better than RNN models for depression detection.

Table 3 reports the generalization ability of our approach on the unseen dataset (Section 3). The models trained using our optimized embedding managed to maintain their performance with generalization ability. Our embedding performs better because it is optimized using the CLPsych2015 dataset, which includes depression and PTSD labeled data. Table 3 shows that the results of the CNN models are competitive, as opposed to RNN models. The best performing RNN model reported 91.425%. CBOW embedding performed the least as compared to others, including the random embedding. In particular, pre-trained CBOW and skip-gram models do not perform as expected, mainly due to the size of the CLPsych2015 corpus, which is nearly around 22 million words. Furthermore, optimized and trainable random embeddings have an advantage for being able to update their weights during training. We conclude that user-level classification for depression detection performs well even with datasets that are small and/or imbalanced.

8 Comparison to Related Work

Resnik et al. (2015) and Preotiuc-Pietro et al. (2015b) reported high results for the CLPsych2015 shared task using topic models. However, their results are not comparable, as they are reported on the official testing set that was not available to us. Alternatively, we performed a five-fold cross-validation on the shared task training data (Tables 2 and 3). We report better performance when testing on the Bell Let's Talk dataset as compared to Jamil et al. (2017).

9 Conclusion

In conclusion, we presented a novel approach to optimize word-embedding for classification tasks. We performed a comparative evaluation on some of the widely used deep learning models for depression detection from tweets on the user level.

We performed our experiments on two publicly available datasets, CLPsych2015 and Bell Lets Talk. Our experiments showed that our CNN-based models perform better than RNN-based models. Models with optimized embeddings managed to maintain performance with the generalization ability.

For future work, we will evaluate against more RNN-based models, in particular with more focus on attention mechanisms. We will investigate other kinds of mental disorders, such as PTSD.

Acknowledgments

This research is funded by Natural Sciences and Engineering Research Council of Canada (NSERC) and the Ontario Centres of Excellence (OCE).

References

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders (5th ed.)*, 5 edition. American Psychiatric Publishing, Washington.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations*, pages 1–15.

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.

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

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, New York. ACM.

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.

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. 2015. Social Media for Mental Illness Risk Assessment , Prevention and Support. In *1st ACM Workshop on Social Media World Sensors*, page 2806659, Guzelyurt.

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

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015)*, Denver.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning Distributed Representations of Sentences from Unlabelled Data. pages 1367–1377.

Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. 2017. Monitoring Tweets for Depression to Detect At-risk Users. pages 32–40.

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.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A Convolutional Neural Network for Modelling Sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 655–665, Baltimore, Maryland, USA. Association for Computational Linguistics.

Sunghwan Mac Kim, Yufei Wang, and Stephen Wan. 2016. Data61-CSIRO systems at the CLPsych 2016 Shared Task. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psycholog*, volume 1, pages 128–132, San Diego, CA, USA. Association for Computational Linguistics.

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.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, pages 1–12.

David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. CLPsych 2016 Shared Task : Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psycholog*, pages 118–127, San Diego, CA, USA. Association for Computational Linguistics.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the Language of Schizophrenia in Social Media. In *Computational Linguistics and Clinical Psychology*, pages 11–20, Colorado. Association for Computational Linguistics.

Andrew. Ng. 2004. Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In *Twenty-first international conference on Machine learning - ICML '04*, page 78, Banff, Alberta, Canada. ACM.

Minsu Park, Chiyoung Cha, and Meeyoung Cha. 2012. Depressive Moods of Users Portrayed in Twitter. In *ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)*, pages 1–8.

James W Pennebaker. 2011. *The secret life of pronouns : what our words say about us*. Bloomsbury Press.

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.

Daniel Preotiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. 2015a.

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. 2015b. 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. 2015. The University of Maryland CLPsych 2015 Shared Task System. In *CLPsych 2015 Shared Task System*, c, pages 54–60.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Nips, Long Beach, CA, USA. Neural Information Processing Systems Foundation.

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

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.