Detection of Depression in Thai Social Media Messages using Deep Learning

Boriharn Kumnunt and Ohm Sornil
Graduate School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand

Keywords: Natural Language Processing, Text Classification, Neural Networks, Depression, CNN, LSTM.

Abstract: Depression problems can severely affect not only personal health, but also society. There is evidence that shows people who suffer from depression problems tend to express their feelings and seek help via online posts on online platforms. This study is conducted to apply Natural Language Processing (NLP) with messages associated with depression problems. Feature extractions, machine learning, and neural network models are applied to carry out the detection. The CNN-LSTM model, a unified model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM), is used sequentially and in parallel as branches to compare the outcomes with baseline models. In addition, different types of activation functions are applied in the CNN layer to compare the results. In this study, the CNN-LSTM models show improvement over the classical machine learning method. However, there is a slight improvement among the CNN-LSTM models. The three-branch CNN-LSTM model with the Rectified Linear Unit (ReLU) activation function is capable of achieving the F1-score of 83.1%.

1 INTRODUCTION

Depression is considered a major problem that affects not only people's well-being, but also their families and society in general. According to statistics from the World Health Organization (WHO), the number of patients with depression problems has increased drastically in the past few years (WHO, 2016). It is globally estimated to affect nearly one in ten people; at its worst, depression can lead to suicide. Besides mental and physical consequences, it can also cause other adverse effects such as reduced work productivity. Further, it requires a significant amount of effort, budget, and time in order to cope with issues (Woo et al., 2011).

Currently, the Internet is an open channel for people to connect and express their thoughts and feelings. In the social media era, there are many methods used for communication between people that have the same interests, such as social media groups and online forums. There is evidence that people with mental illness use social media or internet forums to express their feelings and seek help via posts (Aladağ, Muderrisoglu, Akbas, Zahmacioglu, and Bingol, 2018). This is a major challenge for organizations that provide help to people with symptoms. Interventions with people who have problems can benefit both the individual and the community. By the way, the number of Internet users is enormous. Thus far, the number of internet users is approximately 3.8 billion, a number that exceeds more than half of the world's population (Meeker, 2019). The need for help can arise at any time. It might be a challenge to deal with the number of Internet users. Thus, a system that can perform the initial task of detecting and assessing depression-related messages might be an alternative solution to reduce the prevalence of problems. Previous studies have shown that Natural Language Processing (NLP) could be used effectively to detect and classify messages that are related to depression or other mental health conditions (Yates, Cohan, and Goharian, 2017; Cohan et al., 2018).

The aim of this study is to propose NLP methods as an initial means of detection that can be used to detect whether or not online messages might be associated with depression. In case the result is a classification of depression, useful information or recommendation can be sent to the owner so they can seek help to cope with the problem, such as at a hospital or clinic that they can contact. Moreover, the result can be further used to notify a corresponding organization to provide help in case of an emergency, such as attempting suicide. However, the diagnosis of

Kumnunt, B. and Sornil, O.
Detection of Depression in Thai Social Media Messages using Deep Learning.
DOI: 10.5220/0009970501110118
In *Proceedings of the 1st International Conference on Deep Learning Theory and Applications (DeLTA 2020)*, pages 111-118
ISBN: 978-989-758-441-1

depression for each person requires psychologists or doctors in this field to thoroughly confirm the illness.

In this study, a number of NLP models and configurations of parameters are used for detecting messages associated with depression problems. The methods used for performing detection consist of feature extraction techniques, machine learning, and deep learning models. Recently, a modification structure for deep learning models can be found in a number of researches. An interesting combination is between Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) according to the effective performance received. Thus, this study will compare CNN-LSTM models with classical machine learning models. Another purpose of this study is to compare the results obtained from various types of activation functions used in deep neural network algorithms. All selected methods will be applied in the context of the Thai language with a created dataset from an online forum platform.

2 RELATED WORK

Classical machine learning models, such as logistic regression and Support Vector Machine (SVM) are usually applied for classification task (Dreiseitl and Ohno-Machado, 2002; Tong and Koller, 2001). Another approach for performing the task is the application of Artificial Neural Networks (ANN) such as Recurrent Neural Networks (RNN) and CNN. The RNN has the ability to preserve sequence information along the process (Liu, Qiu, and Huang, 2016). However, if a sequence is too long, it may suffer from short-term memory. Thus, LSTM which is composed of input gate, output gate, and forget gate, is designed to solve the problem (Hochreiter and Schmidhuber, 1997; Gers, Schmidhuber, and Cummins, 1999). Furthermore, the Bidirectional Long Short-Term Memory Networks (BLSTM) is based on the concept of LSTM with the layer that can obtain information from past (backward) and future (forward) states (Tai, Socher, and Manning, 2015). For the CNN, it is originally developed for the task of image processing (LeCun et al., 1989; Cireşan, Meier, Gambardella, and Schmidhuber, 2011). It uses a series of convolutional layers to apply with local features (Kim, 2014).

Recently, some researches have shown the performance of combining CNN with LSTM, and some pieces of research are conducted to propose suitable parameters for the model. Wang, Yu, Lai, and Zhang (2016) used regional CNN-LSTM Model for dimensional sentiment analysis, Yenter and Verma (2017) applied the models in parallel as branches for review sentiment analysis, and Heikal, Torki and El-Makky (2018) used an ensemble model between the CNN and LSTM model for sentiment analysis of Arabic messages. In these studies, the mentioned models return effective outcomes in performing their tasks.

Technically, the neural network architecture is composed of input, output, and hidden layers. The activation functions are used in some layers to determine the model output. They are based on mathematical equations, of which there are various types of functions to be applied in each layer. Comparisons of activation functions have been conducted in some researches (Nam, Kim, Mencía, Gurevych and Fürnkranz, 2014). Each activation function has unique characteristics, advantages, and drawbacks (Nwankpa, Ijomah, Gachagan and Marshall, 2018). The Sigmoid function is claimed to be useful for binary classification problems. The Hyperbolic Tangent Function (Tanh) returns better outcomes for multi-layer neural networks. However, both functions are suffered from vanishing gradient problem, which can be solved by the Rectified Linear Unit (ReLU) function (Nair and Hinton, 2010).

Various researches applied NLP with the purpose of assessing texts associated with depression and other mental disorders. Dataset used for studying in this field cannot be publicly published due to ethical and privacy concerns (Benton, Coppersmith and Dredze, 2017). Thus, some studies had to construct their dataset for performing the classification. Yates et al. (2017) constructed the Reddit Self-reported Depression Diagnosis (RSDD) dataset from www.reddit.com, an online forum platform. The dataset is assessed by two proposed CNN models comparing with classical machine learning methods. Cohan et al. (2018) created the Self-reported Mental Health Diagnoses (SMHD) dataset. Besides self-reported messages with depression, the SMHD dataset also contains the other eight mental health conditions.

In the context of the Thai language, NLP methods have been studied and applied in various fields. Jotikabukkana, Sornlertlamvanich, Manabu and Haruechaiyasak (2015) performed text classification with twitter messages, using term frequency-inverse document frequency (tf-idf) and Word Article Matrix (WAM) techniques. Seneewong Na Ayutthaya and Pasupa (2018) performed Thai sentiment analysis with the application of features extraction, embedding vectors, Part-of-Speech (POS) embedding, sentic features, and a concatenated Bidirectional LSTM-CNN model. However, at the time writing this study,

there is no study applied text classification with depression problems in the context of the Thai language.

3 DATASET

In order to create a dataset of messages associated with depression problems, the collection process is conducted following the construction of the RSDD dataset by Yates et al. (2017). The difference between the RSDD dataset and the collected dataset in this study is that the RSDD dataset is mainly focused on users with self-reported depression diagnoses, while the dataset used in this study is mainly focused on all kinds of text related to problems with depression, such as questions about depression, asking for help, consequences from the illness, and personal experience. Messages are collected from posts on www.pantip.com, an online forum platform in the Thai language. The online forum contains 38 major virtual rooms such as health, music, finance, and travel rooms. Users can read and post information, questions, reviews, and polls on the forum. Some provided hashtags can be chosen and used as labels for posting. The maximum allowable number of hashtags is five hashtags per post. Messages about depression problems are usually posted in the health room and labeled with the depression hashtag.

Firstly, posts from users who used the depression hashtag between January 2017 and February 2020 was collected. From each user, only a post with the depression hashtag was selected. To confirm that the messages in each post were associated with depression, each one had to be labeled with the depression hashtag by the owner of each post. Besides the hashtag, it had to contain words or context related to the conditions of depression (e.g., anxiety, or suicidal). To clean the dataset, posts with editing history or modifications (e.g., edited posts or posts that some paragraphs were deleted after posting.), and posts containing no text (e.g., posts with only pictures or video) were eliminated from the dataset.

For classification, a control dataset with posts that contain no depression hashtag is created. The selection of control users is based on their posting activities on the forum. From 38 rooms, each room on the forum is considered a stratified group for sampling. After the sampling process in each room, the selected control users with the history of using depression hashtag and used to post in the health room are filtered out. Only users who posted within the same period as previously mentioned are selected. Latent Dirichlet Allocation (LDA) model (Rus, Niraula and Banjade, 2014) is employed for matching between the users that used the depression hashtag and control users based on the probability of hashtags used, ignoring the use of the depression hashtag. Then, the minimum Hellinger distance between users is measured to select users with the closest distance. Hellinger distance is used for selecting users that have the same probability in using each hashtag for posting on the forum and preventing bias (Yates et al., 2017). Finally, with the smallest Hellinger distance obtained between users, five control users are selected to be matched with their closest user that used the depression hashtag. The same criteria for cleansing the dataset are applied to the control dataset.

The obtained dataset contains 5,283 posts with depression hashtags and 26,415 control posts. In total, the dataset contains 31,698 posts. Table 1 provides an overview of the dataset with descriptive statistics, average, median, standard deviations, and the total number of posts grouped by type of hashtag. The table shows information in the tokenization and character level. It can be observed that, even though the opposed hashtag outnumbered the number of posts with the depression hashtag, the average of tokened words per post with the depression hashtag was higher.

4 METHODOLOGY

The process of processing and constructing the models will be explained in this section. The first step is pre-processing with tokenization and removing stopwords, followed by feature extraction with tf-idf and word embedding methods. Then, classical machine learning and deep learning models will be applied for performing text classification. The deep learning techniques will be adjusted with different types of structure and activation functions. Lastly, the model structures will be shown in figure 1, and their parameters will be summarized in table 2.

Table 1: Statistics overview of the dataset in tokenization level and characters level in parenthesis.

Hashtag	Descriptive statistics			
	Mean	Median	Std. Dev.	Total
Depression	119.0 (572.3)	74 (363)	139.8 (667.2)	628.7k (3.02m)
Non Depression	71.2 (328.9)	35 (161)	114.6 (533.5)	1,881.8k (8.69m)
All	79.2 (369.5)	39 (181)	120.5 (565.3)	2,510.5k (11.71m)

4.1 Pre-processing

Linguistically, there are no spaces used between words in each sentence for writing in the Thai language, unlike in English. The first step is to prepare words for the processing; tokenization is applied for each sentence of the dataset. PyThaiNLP, which is a Python package for text processing, is used for word segmentation. The tokenization algorithm used in this study is "newmm", which is a dictionary-based tokenizer. The version used is 2.1.4. The last step in the pre-processing is removing stop words, which are words that are grammatically necessary, but do not have any meaning about the text content from the sentences.

4.2 Feature Extraction

Feature extraction is theoretically considered a useful and practical pre-processing step that not only conducts dimensionality reduction for removing irrelevant and redundant data, but also increases learning accuracy and improves outcome accuracy (Khalid, Khalil and Nasreen, 2014). Term frequency-inverse document frequency (tf-idf) weighting is the first technique used in this study. The weight obtained from the tf-idf weighting is a statistical measure that helps to avoid the problem of words that appear more frequently become the dominator of the dataset (Waykole and Thakare, 2018).

Word embedding is another approach for extracting features. The Word2Vec algorithms are used to generate word vectors from the dataset. The algorithm is capable of transforming each word in the sentence into a vector positioned in vector space (Mikolov, Chen, Corrado and Dean, 2013). Words that share similar contexts in the corpus will be located close to one another in the vector space. Gensim, which is a library for NLP on Python, is used for creating Word2vec embedding from the collected dataset. The version of Gensim used is 3.8.3.

4.3 Depression Classification

The models for the detection of messages associated with depression problems consist of classical machine learning methods and deep neural network models. For the classical approach, the first classification model is Logistic regression, which is a useful model for performing classification (Dreiseitl and Ohno-Machado, 2002). It uses a logistic function to model a binary dependent class. Another model is Support Vector Machine (SVM), which is commonly applied for detection (Tong and Koller, 2001). The SVM model uses a separating hyperplane as a discriminative classifier. Both logistic regression and SVM models will be used with the feature extraction techniques to create the baseline models in this study.

For the deep learning models, the plain LSTM (Hochreiter and Schmidhuber, 1997; Gers et al., 1999) will be applied as another baseline model. The

Table 2: Parameters setup for the deep learning models.

Parameters		Architecture		
		LSTM	CNN-LSTM	Three-branch CNN-LSTM
CNN	Kernel sizes	-	5	3 / 4 / 5
	Filters	-	50	32
	Activation functions	-	ReLU / Sigmoid / Tanh	ReLU / Sigmoid / Tanh
	Max Pooling	-	2	2
	Dropout	-	0.1	0.1
LSTM	Units	64	64	32
	Activation functions	Tanh	Tanh	Tanh
Dense	Units	2	2	2
	Activation functions	Sigmoid	Sigmoid	Sigmoid
Optimizer	Type	Adam	Adam	Adam
	Learning Rate	0.001	0.001	0.001

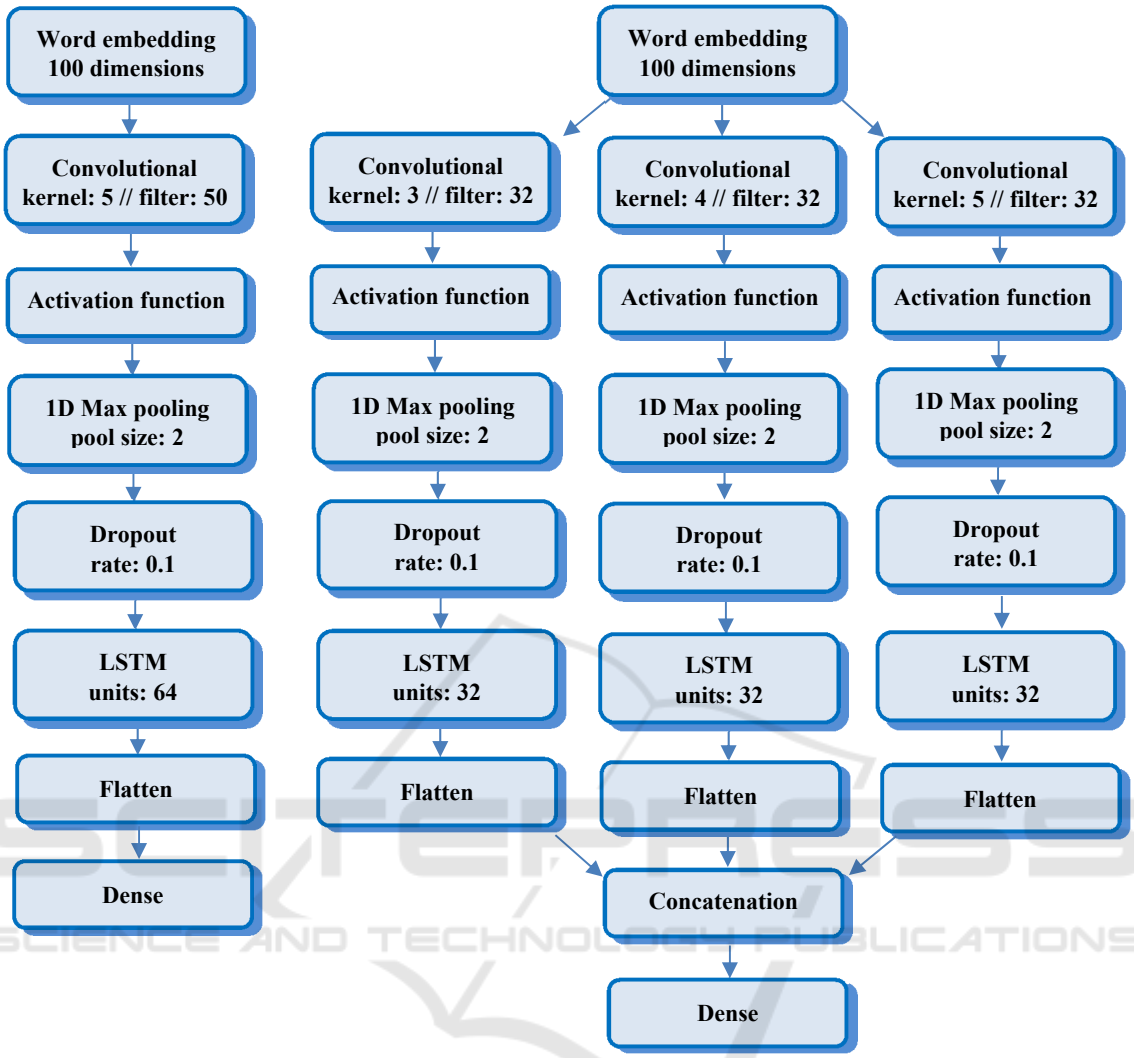

Figure 1: Comparison of model diagrams between the CNN-LSTM model (left) and the three-branch CNN-LSTM model (right).

CNN-LSTM models (Wang et al., 2016; Yenter and Verma, 2017), which are created by combining the CNN (LeCun et al., 1989) and LSTM (Liu et al., 2016, Hochreiter and Schmidhuber, 1997) layers will be used in a number of different architectures, together with three different types of activation functions in the CNN layer, i.e., the Sigmoid, Tanh, and ReLU functions. For the model construction, the convolutional and LSTM layer in the CNN-LSTM models will be constructed sequentially and in parallel to compare the results between them and the baseline models.

4.4 Model Architectures

The plain LSTM model will be used without combining with other layers except flatten and dense layer for returning the classification result. The CNN-LSTM architectures are based on the models used in the study of Wang et al. (2016) and Yenter and Verma (2017). The principal model structure is shown on the left side of Figure 1. It consists mainly of convolutional, pooling, dropout, LSTM, and dense layers. After creating sequences of words, firstly, the convolutional layer will receive and process the data with its activation function. Then the output will be sent to the pooling and dropout layer, respectively, followed by the LSTM layer, which is set up with the Tanh activation function. In the dense layer, the

Softmax activation is used to return the output. The dense layer uses stochastic gradient descent with Adam optimizer (Kingma and Ba, 2015). All parameters are set up as shown in Table 2.

5 EXPERIMENTS

5.1 Experimental Setup

Due to the different proportions of classes within the dataset, data balancing technique is performed to cope with the issue. Firstly, for solving the minority class problem, training and test sets are chosen by using stratified 10-fold cross-validation to avoid overlapping between the training and test sets. Another purpose is to preserve the percentage of samples in each class. Furthermore, the class weighting technique, based on the study by King and Zeng (2001), is applied as a balancing technique in each fold of cross-validation.

5.2 Parameters Setup

The parameters for Word2vec are varied by tuning embedding size (Yates et al., 2017) between 100, 120, 150, 200, 250, and 300, and learning rate between 0.01, 0.05, 0.1, and 0.5. For tuning the tf-idf scheme, the maximum feature size is set at 100, 120, 150, 200, 250, and 300. The Word2Vec scheme is based on the Continuous Bag-of-Words (CBOW) model. In the detection process with the machine learning model, Logistic regression and SVM models are L2 normalized. For the plain LSTM model, the number of units set up is 64. The CNN layer in the CNN-LSTM model with no branches is set up with 50 filters and a kernels size of 5. Three different types of activation functions, which are the Sigmoid, Tanh, and ReLU activation functions, are applied in the CNN layer in each model to compare the outcomes.

In order to extend the aforementioned CNN-LSTM models, modified model structures are constructed by laying layers in parallel as branches instead of placing them sequentially. The number of branches applied in this study is three branches (Yenter and Verma, 2017), as shown on the right side of Figure 1. For the three-branch CNN-LSTM models, the number of filters in the CNN layer is 32. The number of units in LSTM is 32. The kernel sizes in the CNN layer are tested between 2 and 5. For the LSTM layer, the number of units is set up between 32 and 64. The final parameters used in this study are explained in Table 2. For the validation process, all models are performed with the same stratified 10-fold cross-validation. The number of the batch size is 100. Lastly, the early stopping is processed by determining the trend of the classification accuracy to prevent overfitting.

5.3 Results

Table 3 shows results obtained from all classification models applied with the dataset, which contains messages associated with depression and non-depression. In the table, precision, recall, and F1-score from the experiment are shown. The baseline models consist of the Logistic regression and SVM models used with features extracted by the tf-idf and Word2Vec algorithms. The deep neural network models are the LSTM model, CNN-LSTM models, and three-branch CNN-LSTM models. In the CNN layer of each model, different types of activation functions are used, one activation function at a time.

According to the results in Table 3, we can see that the highest F1-score from the four baseline models with classical machine learning models are 0.576, 0.588, 0.648 and 0.770. The highest score is archived from the SVM model with features from the Word2Vec algorithm. These F1-score results that are near statistical (50%) might be due to the set of parameters such as maximum feature or learning rate used in this study being limited to optimize the models. However, the highest score of 0.770 is quite close to the results from the deep neural network models.

The plain LSTM model, which is another model used as baseline model, returns the F1-score of 0.823. The result from the LSTM model is a bit lower than the results from the ensemble CNN-LSTM models. For the CNN-LSTM and three-branch CNN-LSTM models, an increasing number of branches in the

Table 3: Depression detection results.

Models [a]		Precision	Recall	F1
Tf-idf LR		0.681	0.503	0.576
Tf-idf SVM		0.640	0.550	0.588
Word2vec LR		0.570	0.869	0.648
Word2vec SVM		0.753	0.832	0.770
LSTM		0.898	0.761	0.823
CNN-LSTM	Sigmoid	0.897	0.762	0.824
	Tanh	0.908	0.753	0.824
	ReLU	0.910	0.760	0.827
Three-branch CNN-LSTM	Sigmoid	0.896	0.769	0.826
	Tanh	0.901	0.769	0.829
	ReLU	0.899	0.772	**0.831**

model structures slightly improves the F1-score with 0.2%, 0.5%, and 0.4% in the models with the Sigmoid, Tanh and ReLU activation functions respectively. The highest F1-score is 83.1% received from the three-branch CNN-LSTM model with the ReLU activation function. The first runner-up is the three-branch CNN-LSTM model with the Tanh activation function with an F1-score of 82.9%. From the empirical outcomes in this study, increasing the number of branches seems to slightly improve the classification results. This might be due to the limited number of parameters used for tuning.

The higher F1-score from the models with the ReLU activation function (Nair and Hinton, 2010) can be explained by the function having the capacity to disable negative activation, so it reduces the number of parameters to be learned (Nam et al., 2014) and returns the improved performance. The outcomes in this study are in accordance with the discussion in the study of Nwankpa et al. (2018) that the ReLU function gives better results compared with the Sigmoid and Tanh activation functions in deep learning application.

6 CONCLUSION AND FUTURE WORK

The initial detection and assessment of messages associated with depression problems might be an alternative solution to helping people cope with their problems. Thus, NLP techniques are applied to classify social media messages associated with depression. In order to improve and produce accurate results, various models are applied to execute the task. In this study, the models with the ReLU activation function yield better results compared with the baseline models and the models with the Sigmoid and Tanh activation function. From the results, the deep learning models did not return a significant improvement from each other, which might be due to the parameters used not being varied enough to optimize the models.

For further works, a wide range of parameters could be tuned to optimize the models. The detection models can be studied further by applying different kinds of activations functions (e.g., PReLU or Softplus activation function). More layers of the CNN or LSTM could be added to extend the models. Furthermore, changing or combining different types of architectures (e.g., Recurrent Neural Network (RNN), or Gated Recurrent Unit (GRU)) could also be explored. Similar architectures and configurations could be applied to other tasks, such as sentimental analysis and topic segmentation.

ACKNOWLEDGEMENTS

This study is supported by the faculty of the Graduate School of Applied Statistics (GSAS), National Institute of Development Administration (NIDA), Bangkok, Thailand.

REFERENCES

World Health Organization, 2016. World health statistics 2016: monitoring health for the SDGs, sustainable development goals, *WHO Press. Switzerland*.

Woo, J. M., Kim, W., Hwang, T. T., Frick, K. D., Choi, B. H., Seo, Y. J., Kang, E. H., Kim, S. J., Ham, B. J., Lee J. S. & Park Y. L., 2011. Impact of Depression on Work Productivity and Its Improvement after Outpatient Treatment with Antidepressants. *Value in Health,* 14(4):475–482.

Aladağ, A. E., Muderrisoglu, S., Akbas, N. B., Zahmacioglu, O. & Bingol, H. O., 2018. Detecting suicidal ideation on forums: proof-of-concept study. *Journal of medical Internet research,* 20(6):e215.

Meeker, M., 2019. Internet Trends 2019. *Code Conference 2019.* [online] Available at: https://www.bondcap.com/pdf/Internet_Trends_2019.pdf [Accessed 26 Apr. 2020].

Yates, A., Cohan, A. & Goharian, N., 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* 2958–2968.

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S. & Goharian, N., 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,* 1485– 1497.

Dreiseitl, S. & Ohno-Machado, L., 2002. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics, 35*(5-6), 352-359.

Tong, S. & Koller, D., 2001. Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research,* 2:45-66.

Liu, P., Qiu, X. & Huang, X., 2016. Recurrent neural network for text classification with multi-task learning. In *Proceedings of International Joint Conference on Artificial Intelligence,* 2873–2879.

Hochreiter, S. & Schmidhuber, J., 1997. Long short-term memory. *Neural computation,* 9(8):1735–1780.

Gers, F. A., Schmidhuber, J. & Cummins, F., 1999. Learning to forget: Continual prediction with lstm. *Neural Computation,* 12:2451–2471.

Tai, K. S., Socher, R. & Manning, C. D., 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,* 1556-1566.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E. & Jackel, L. D., 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation,* 1(4): 541–551.

Cireşan, D. C., Meier, U., Gambardella, L. M. & Schmidhuber, J., 2011. Convolutional neural network committees for handwritten character classification. In *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR),* 1135–1139.

Kim, Y., 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNL), October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL,* 1746–1751.

Wang, J., Yu, L. C., Lai, K. R. & Zhang, X., 2016. Dimensional sentiment analysis using a regional CNNLSTM model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* 225-230.

Yenter, A. & Verma, A., 2017. Deep cnn-lstm with combined kernels from multiple branches for imdb review sentiment analysis. In *Proceedings of the 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference,* 540–546.

Heikal, M., Torki, M. & El-Makky, N., 2018. Sentiment Analysis of Arabic Tweets using Deep Learning. *In Proceedings of the International Conference of procedia Computer Science,* 142:114-122.

Nam, J., Kim, J., Mencía, E. L., Gurevych, I. & Furnkranz, J., 2014. Large-scale multi-label text classification - revisiting neural networks. In *Proceedings of Machine Learning and Knowledge Discovery in Databases. European Conference,* 437–452.

Nwankpa, C., Ijomah, W., Gachagan, A. & Marshall, S., 2018. Activation functions: Comparison of trends in practice and research for deep learning. *Cornell University.* arXiv preprint arXiv:1811.03378.

Nair, V. & Hinton, G. E., 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the International Conference on Machine Learning,* 807–814.

Benton, A., Coppersmith, G. & Dredze, M., 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing,* 94–102.

Jotikabukkana, P., Sornlertlamvanich, V., Manabu, O. & Haruechaiyasak, C., 2015. Effectiveness of social media text classification by utilizing the online news category. In *Proceedings of the 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications,* 1-5.

Ayutthaya T. S. N. & Pasupa, K., 2018. Thai Sentiment Analysis via Bidirectional LSTM-CNN Model with Embedding Vectors and Sentic Features. In *Proceedings of the International Joint Symposium on Artificial Intelligence and Natural Language Processing,* 1-6.

Rus, V., Niraula, N. & Banjade, R., 2013. Similarity measures based on latent Dirichlet allocation. *Computational Linguistics and Intelligent Text Processing,* Gelbukh A., Springer, 459–470.

Khalid, S., Khalil, T., & Nasreen, S., 2014. A survey of feature selection and feature extraction techniques in machine learning. In *Proceedings of 2014 Science and Information Conference (SAI),* IEEE, 372–378.

Waykole, R., & Thakare, A., 2018. A Review of Feature Extraction Methods for Text Classification. *International Journal of Advance Engineering and Research Development,* 5(4):351-354.

Mikolov, T., Chen, K., Corrado, G. & Dean, J., 2013. Efficient Estimation of Word Representation in Vector Space. In *Proceedings of Workshop at International Conference on Learning Representations*. [online] Available at: http://arxiv.org/pdf/1301.3781.pdf [Accessed 26 Apr. 2020].

Kingma, D., & Ba, J., 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations,* 1-15.

King, G. & Zeng, L., 2001. Logistic regression in rare events data. *Political Anal,* 9(2):137–163.