Did you take the pill? - Stacked Ensemble of CNNs for Identifying Mentions of Personal Intake of Medicine in Twitter

Jasper Friedrichs
Lyft, Palo Alto, USA
Email: jasper.friedrichs@gmail.com

Debanjan Mahata
Bloomberg L.P., New York, USA
Email: dmahata@bloomberg.net

Rajiv Ratn Shah
Singapore Management University, Singapore
Email: rajivshah@smu.edu.sg

Jing Jiang
Singapore Management University, Singapore
Emai: jingjiang@smu.edu.sg

***Abstract*—Mining social media messages such as tweets, articles, and Facebook posts for health and drug related information has received significant interest in pharmacovigilance research. Social media sites (e.g., Twitter), have been used for monitoring drug abuse, adverse reactions of drug usage and analyzing expression of sentiments related to drugs. Most of these studies are based on aggregated results from a large population rather than specific sets of individuals. In order to conduct studies at an individual level or specific cohorts, identifying posts mentioning intake of medicine by the user is necessary. Towards this objective we develop a classifier for identifying mentions of personal intake of medicine in tweets. We train a stacked ensemble of shallow convolutional neural network (CNN) models on an annotated dataset. We use random search for tuning the hyper-parameters of the CNN models and present an ensemble of best models for the prediction task. Our system produces state-of-the-art result, with a micro-averaged F-score of 0.693.**

I. Introduction

Social media has become an ubiquitous source of information for a variety of topics. Right from information related to daily events, personal rants, to expressions of intake of medicine and adverse drug reactions, are readily available in publicly accessible social media channels such as Twitter[1], DailyStrength[2], MedHelp[3], among others. Huge amounts of data made available on these platforms have become an useful resource for conducting public health monitoring and surveillance, commonly known as pharmacovigilance [1]. The work presented in this paper aims at identifying intake of personal medication expressed by an user in Twitter. The broader perspective of such a system is to aid in developing automated methods for performing pharmacovigilance activities in social media, and to study the effects of medicine on an individual as well as specific cohorts [2].

Substantial attempts have been made to mine social media content in order to identify adverse drug reactions [3], abuse [4], and user sentiment [5], from posts mentioning medications. However, all these studies are based on aggregated results from large set of content that mentions a medicine/drug, without taking into account whether the user has actually consumed the medicine/drug. Without this knowledge, a true assessment of the effects of medication intake in general and how it affects a specific group of users cannot be done. In order to leverage social media data for studying targeted groups and to do a true assessment of the effects of medication intake, it is necessary to develop systems that can automatically distinguish posts that expresses personal intake of medicine from those that do not. In this work we only concentrate on Twitter as the social media channel for performing such a task.

The key to the process of identifying tweets mentioning personal intake of medicine and to draw insights from them is to build accurate text classification systems. The effectiveness of developing classifiers has already been shown to be useful in identifying adverse drug reactions expressed in Twitter [3]. However, mining social media posts comes with unique challenges. Microblogging websites like Twitter pose challenges for automated information mining tools and techniques due to their brevity, noisiness, idiosyncratic language, unusual structure and ambiguous representation of discourse. Information extraction tasks using state-of-the-art natural language processing techniques, often give poor results for tweets. Abundance of link farms, unwanted promotional posts, and nepotistic relationships between content creates additional challenges [6].

The main objective of the task presented in this paper is to categorize short colloquial tweets into one of the three categories,

1) ***personal medication intake*** *(Class 1)* - tweets in which the user clearly expresses a personal medication intake/consumption (e.g. *I had the worst headache ever and I just took an AdvilRelief #advil and now I feel so much better thank*).
2) ***possible medication intake*** *(Class 2)* - tweets that are ambiguous but suggest that the user may have taken the

[1]http://twitter.com
[2]https://www.dailystrength.org/
[3]http://www.medhelp.org/

medication (e.g. *I should have taken advil on friday then i might have actully had an amazing weekend.. instead of throwing up 20 times a day #advil, not this time*)

3) ***non-intake*** *(Class 3)* - tweets that mention medication names but do not indicate personal intake (e.g. *Understand the causes and managing #Migraine Madness #aspirin #diet #botox #advil #relpax #headache*).

Towards the above goal, we design and implement a deep learning classifier - *Stacked Ensemble of Shallow Convolutional Neural Networks* (Section III-A), trained on an annotated dataset provided at SMM4H-2017 shared task workshop[4]. We compare the results of our classification system with other classifiers that participate in the shared task and get state-of-the-art results, with a micro-averaged F-score of 0.693 for Class 1 and 2. We submitted our system (*InfyNLP*) at the workshop and was ranked first amongst 26 submissions [7], [8]. In this paper, we intend to elaborately discuss and present our submitted system as well as our model choices and learnings.

Next, we present work related and relevant to the scope of this paper.

II. Related Work

Pharmacovigilance is the branch of pharmacological sciences that deals with drug safety. It studies the collection, detection, monitoring, assessment, and prevention of harmful effects with pharmaceutical products. Mining social media messages such as tweets, blog articles, and Facebook posts for health and drug related information has received significant interest in pharmacovigilance research as people tend to share their daily activities on social media. Since pharmacovigilance heavily focuses on adverse drug reactions, it necessitates an automatic detection of the personal intake of medicine by sensing the social media to build smart health-care applications. Thus, developing automated classification models for identifying messages (e.g., tweets) containing description of personal intake of medicine is a pragmatic step towards the automation of Pharmacovigilance. Cramer *et al.* [9] found that as we increase the number of dosages of medicines (say from one to four times) in a day, compliance rate decreases. Thus, several research work [1], [10], [11] discussed and explored different healthcare problems through advanced technologies. In this section, we provide a brief literature review on this problem.

Due to active participation of users on social media, it is now feasible to classify latent user attributes (*e.g.*, gender, age, regional origin, and political orientation) in social media such as Twitter [21]. Research studies in last few years indicate that social media is heavily used in building healthcare applications [22], [23], [24]. Sarker *et al.* [25] presented a review of pharmacovigilance techniques from social media data and discussed a possible pathway for automated pharmacovigilance research.

[4]https://healthlanguageprocessing.org/sharedtask2/

Deep learning techniques have yielded immense success in computer vision, natural language processing, speech processing, machine translation, and healthcare [12], [13], [14], [15]. However, training deep neural networks to obtain good models is not easy and depends on determining hyper-parameters optimally [16]. Bergstra *et al.* [17] performed random search for hyper-parameter optimization. Moreover, Lim *et al.* [18] used deep neural network techniques to detect user-level psychological stress from social media. However, only going deeper with convolutions does not lead to the best solution [19]. Recent studies such as Le *et al.* [20] explored shallow networks for text classification. They found that their shallow word models outperform deeper models. This encouraged us to use a shallow network in our proposed approach for identifying personal medication intake from Twitter.

Often one solution to a complex problem does not fit to all scenarios [26]. Thus, researchers use ensemble techniques to address such problems. Zhou *et al.* [32] presented a neural network ensemble and proposed, that many neural networks can be jointly used to solve a problem efficiently. For instance, Deng and Platt [33] use an ensemble of deep learning models for speech recognition. Wang *et al.* [27] presented *ensemble of classifier* approaches for biomedical named entity recognition by combining *generalized winnow*, *conditional random fields*, *support vector machine*, and *maximum entropy* through three different strategies. Moreover, ensemble techniques have shown to perform well in biomedical entity extraction [28] and named entity recognition [29], [30]. Furthermore, stacked ensemble techniques are very useful in different healthcare applications [31], [30]. A recent work [34] proposed an ensemble of several classifiers that effectively distinguishes between adverse drug events (ADEs) and non-ADEs from informal text in social media.

Our literature review confirms that leveraging social media data using ensemble and neural network techniques is very beneficial in healthcare applications. Thus, in order to solve the problem of identifying personal medication intake from Twitter, we train a stacked ensemble of shallow convolutional neural network (CNN) models on an annotated dataset. We use random search for tuning the hyper-parameters of the CNN and build an ensemble of best models that achieves state-of-the-art performance.

III. Methodology

Deep learning systems have recently shown to achieve top results in tasks related to natural language processing on tweets [24]. Historically, ensemble learning has proved to be very effective in most of the machine learning tasks including the famous winning solution of the Netflix Prize [26]. Ensemble models can offer diversity over training data splits, random initialization of the same model or model architectures, and a combination of multiple average or low performing learners to produce a robust and high-performing learning model. A convolutional neural network (CNN) is a deep learning architecture, that has shown strong performance on sentence-level text classification [12]. Even fairly simple CNNs evaluate

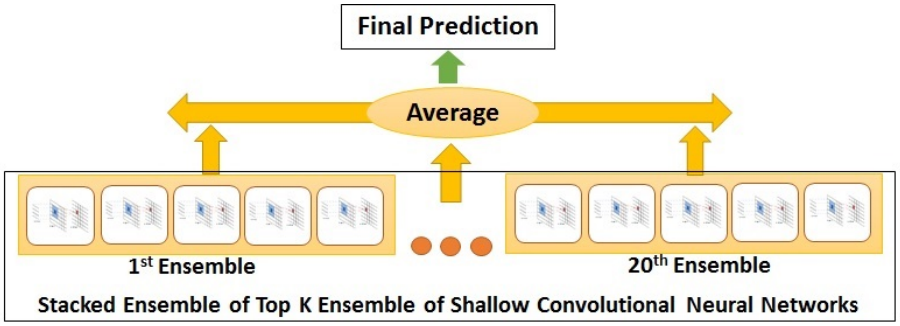

Fig. 1: A Stacked Ensemble of 100 (20 x 5) shallow convolutional neural networks.

at a level of or even better than more complex deep learning architectures [20]. Therefore, we design and implement a stacked ensemble of shallow convolutional neural networks (Figure 1) for solving the classification task presented in this paper. The main intuition behind developing such an ensemble was to take the best of both worlds. Next, we explain the architecture of stacked ensemble of CNNs that we train.

A. Stacked Ensemble of Shallow Convolutional Neural Networks

A Stacked Ensemble of Shallow Convolutional Neural Networks is a large ensemble classifier comprising of smaller ensembles stacked over one another, with the underlying classifier being a standard shallow Convolutional Neural Network (CNN) model similar to that used in [12]. In order to train such an ensemble model we enlist the generic steps:

Step 1 Train a shallow CNN model on each fold while performing *c-fold* cross validation on the training dataset.

Step 2 The output of each model trained on each fold is averaged to get the final output of an ensemble of c CNN models ($ensemble_i^{output}$, Equation 1).

Step 3 Train n such ensembles as in Step 2.

Step 4 Sort the n ensembles in terms of their performance on the metric suitable for the classification task.

Step 5 Choose top K ensembles based on their performance on the training dataset to form the final stacked ensemble of K CNN ensemble models.

Step 5 The final output prediction ($stacked-ensemble^{output}$), is given by the average of the predictions made by each of the top K ensembles (Equation 2).

$$ensemble_i^{output} = average(prediction_1, prediction_2, ..., prediction_c) \quad (1)$$

$$stacked - ensemble^{output} = average(ensemble_{top_1}^{output}, ensemble_{top_2}^{output}, ..., ensemble_{top_K}^{output}), \quad (2)$$

Figure 1, shows a high level architecture of the final stacked ensemble of CNNs that we use in predicting the outcome of the task presented in this paper. We train a standard shallow CNN model, on each fold while performing 5-fold cross validation on our training dataset. We take the output prediction of each of these models trained on each fold and average them to create an ensemble of 5 models. We further train 20 such ensembles. For the final output we sort the ensembles in order of their decreasing performance on the training dataset and take the $top\ k$ ensembles. We take the output of each ensemble and average them to create our stacked ensemble of shallow CNNs. We call this architecture as a stacked ensemble as we stack one ensemble over another in order to create the final ensemble of models that we use for prediction. In general, we can take top K such ensembles and create a stacked ensemble of top K ensemble of shallow CNNs.

In order to get the best results from any classification model, hyperparameter tuning is a key step and CNNs are no exception. While the existing literature offers guidance on practical design decisions, identifying the best hyperparameters of a CNN requires experimentation. This requires evaluating trained models on a cross-validation dataset and choosing the best hyperparameters manually that produce best results. Automated hyperparameter searching methods like *grid search*, *random search*, and *bayesian optimization* methods are also popularly used. In our presented system we use random search [17], to explore the hyperparameters of a shallow CNN architecture and form an ensemble of the best models, which we refer to as a stacked ensemble. Next, we share the detailed settings, output and analysis of our experiment.

IV. Experiment

In this section, we present the experiments that we perform on the dataset in order to achieve the task presented in this paper. We give an overview of the dataset on which we train our models, and discuss about the hyperparameter settings. Results of our experiments are presented accompanied by a discussion of

A. Dataset

The dataset used in this paper is publicly available and can be obtained from 2nd Social Media Mining for Health

Applications Shared Task at AMIA 2017 website[5]. The organizers of the task provided 8000 annotated tweets as a training dataset and 2260 additional tweets as development dataset. We collected the tweets using the script provided along with the dataset, by querying Twitters API. However, we could not collect all the tweets as some of them were not available at the moment when we executed our collection process. Later, the organizers also shared the test dataset, that was used for calculating the final scores of the submitted models. The test dataset consisted 7513 tweets. A distribution of tweets provided for each class and the mapping of each class is shown in Table 1. It is to be noted over here that for training our models, we combine the training and development dataset provided and treat it as our training dataset, therefore learning our models using 9663 tweets with 5-fold cross validation.

	Class 1	Class 2	Class 3	Total
Train	1847	3027	4789	9663
Test	1731	2697	3085	7513

TABLE I: Shared task data distribution. Class 1, 2 and 3 represents *personal medication intake*, *possible medication intake* and *no medication intake*, respectively.

B. Data Preprocessing

We use Spacy[6] for all our data preprocessing and cleaning activities. We do not remove stopwords. Each document in our training and test dataset is converted to a fixed size document of 47 words/tokens. We use two pre-trained word embeddings - godin [35] and shin [14], shared by the authors. Each of these embeddings are of 400 dimensions. Each word in the input tweet is represented by its corresponding embedding vector, when present in the vocabulary of the model.

Hyperparameter	Range
adam_b2	0.9, 0.999
n_dense_output	100, 200, 300, 400
keep_prob (dropout)	0.4, 0.5, 0.6, 0.7, 0.8, 0.9
batch_size	50, 100, 150
learning_rate	0.0001, 0.001
word_embedding	godin [35], shin [14]
n_filters	100, 200, 300, 400
filter_sizes	[1,2,3,4,5], [2,3,4,5,6], [3,4,5,6,7], [1,2,2,2,3], [2,3,3,3,4], [3,4,4,4,5], [4,5,5,5,6]

TABLE II: Hyperparameter ranges used for random search permutations.

C. Hyperparameter Settings for CNNS

We use Xavier weight initialization scheme [16], for initializing the weights of the CNNs. Adam [36] with two annealing restarts has been shown to work faster and perform better than SGD in other NLP tasks [15]. Therefore, we use the same as our optimization algorithm. We use five filters with varying filter sizes in the convolution layer and use dropout during the training process. The models are implemented using TensorFlow[7]. The entire ranges of the hyperparameters that we give to our random search procedure is shown in Table II. The word embedding model to be used during training is also treated as a hyperparameter.

[5]https://healthlanguageprocessing.org/sharedtask2/

[6]https://spacy.io/

Systems	Micro-averaged Precision Class 1 and 2	Micro-averaged Recall Class 1 and 2	Micro-averaged F-score Class 1 and 2
InfyNLP	**0.725**	0.664	**0.693**
UKNLP	0.701	**0.677**	0.689
NRC-Canada	0.704	0.635	0.668
TJIIP	0.654	0.664	0.659
CSaRus-CNN	0.709	0.604	0.652

TABLE IV: My caption

D. Results

$$Precision_{1+2} = \frac{TP_1 + TP_2}{TP_1 + FP_1 + TP_2 + FP_2} \tag{3}$$

$$Recall_{1+2} = \frac{TP_1 + TP_2}{TP_1 + FN_1 + TP_2 + FN_2} \tag{4}$$

$$F - Score_{1+2} = \frac{2 * Precision_{1+2} * Recall_{1+2}}{Precision_{1+2} + Recall_{1+2}} \tag{5}$$

An ensemble of five CNNs is trained during 5-fold cross-validation training performed on our combined training dataset along with random search on the hyperparameter ranges. We train twenty 99 such ensembles. The performance of each the top 20 such ensemble on the training data (blue) and on the test data (red) is shown in Figure 2. The models are arranged in the order of their decreasing training performance. We create stacked ensembles from these ensembles by taking top K ensemble models. We show the performances for such top K stacked ensembles (brown), as well. The detailed performances on the evaluation metrics of Ttop 3, Ttop 10 and Ttop 20 stacked ensembles are shown in Table 3, and denoted by stars in Figure 2. The stacked ensemble formed using top 20 best performing ensembles was submitted to the task, which achieved the best micro averaged F1 score on the tasks test dataset. It can be also observed from Figure 2, that the fifth best ensemble model achieves the best scores on the test dataset. This proves an overall effectiveness of ensemble models in boosting performance on the present classification task.

V. Conclusion and Future Work

By participating in this shared task we showed the generic effectiveness of CNNs and ensembles on identification of personal medication intake from Twitter posts. Our proposed architecture of stacked ensemble of shallow CNNs, outperformed other models submitted in the task. This provided an empirical evaluation of our initial aim of combining ensembles with CNNs along with training the models using

[7]https://www.tensorflow.org/

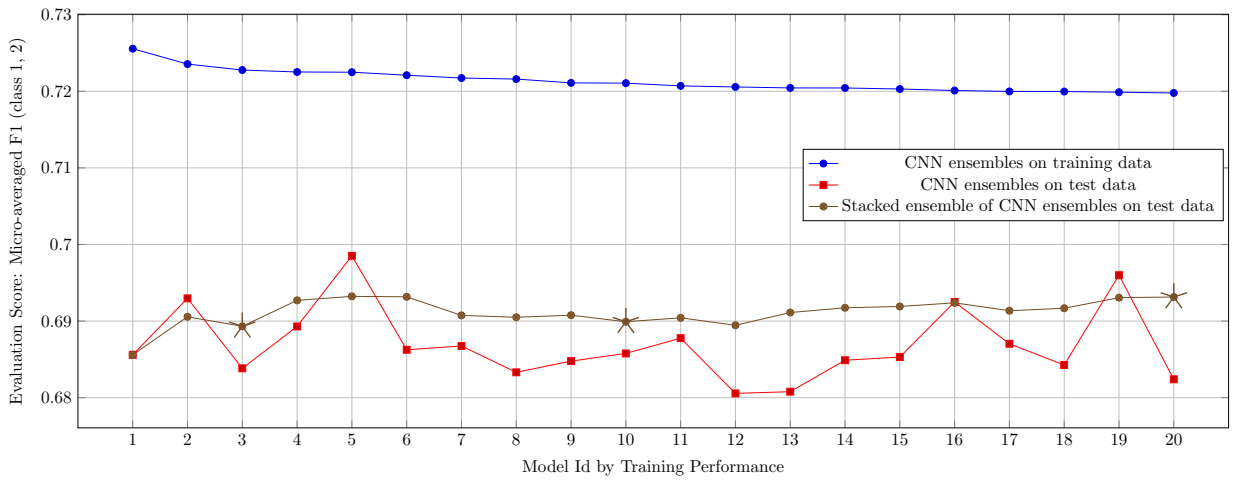

Fig. 2: Individual 5-fold data ensemble and collective parameter ensemble (stacked ensemble) results for top 20 random search models. Models are sorted from left to right by decreasing 5-fold cross validation results.

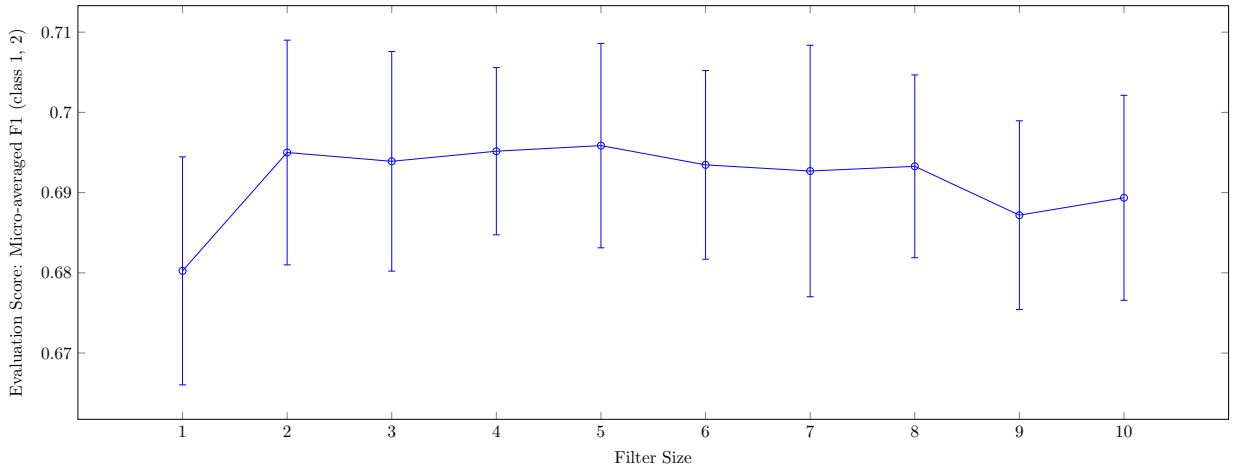

Fig. 3: Individual 5-fold data ensemble and collective parameter ensemble (stacked ensemble) results for top 20 random search models. Models are sorted from left to right by decreasing 5-fold cross validation results.

	Recall			**Precision**			**F1**			**Recall_m**	**Precision_m**	**F1_m**
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**			
top 3	0.696	0.644	0.842	0.704	0.725	**0.763**	0.700	0.682	0.800	**0.664**	0.716	0.689
top 10	0.685	0.646	0.849	0.709	0.729	0.758	0.697	0.685	0.801	0.661	0.721	0.690
top 20	**0.690**	**0.648**	**0.853**	**0.712**	**0.733**	0.761	**0.701**	**0.688**	**0.804**	**0.664**	**0.725**	**0.693**$*$

TABLE III: Evaluation of ensembles on test data. _m stands for micro average recall over class 1 and 2. $*$ marks the state-of-the-art micro averaged F1 on the task's dataset achieved by our best model.

random search on the hyperparameters. In the future, we plan to work more on hyperparameter tuning using random search and various other search procedures and analyze their effectiveness. Instead of using pre-trained word embeddings it would also be interesting to look at the performance of our models by training word and phrase embeddings on a domain specific dataset of tweets. We would also like to formalize the architecture of stacked ensembles of CNNs and compare our models with an exhaustive set of other deep learning as well as traditional machine learning models.

REFERENCES

[1] L. Härmark and A. Van Grootheest, "Pharmacovigilance: methods, recent developments and future perspectives," *European journal of clinical pharmacology*, vol. 64, no. 8, pp. 743–752, 2008.

[2] A. Klein, A. Sarker, M. Rouhizadeh, K. O'Connor, and G. Gonzalez, "Detecting personal medication intake in twitter: An annotated corpus and baseline classification system," *BioNLP 2017*, pp. 136–142, 2017.

[3] A. Nikfarjam, A. Sarker, K. OConnor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *Journal of the American Medical Informatics Association*, vol. 22, no. 3, pp. 671–681, 2015.

[4] C. L. Hanson, S. H. Burton, C. Giraud-Carrier, J. H. West, M. D. Barnes, and B. Hansen, "Tweaking and tweeting: exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students," *Journal of medical Internet research*, vol. 15, no. 4, 2013.

[5] I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, and G. H. Gonzalez, "Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts," *Journal of biomedical informatics*, vol. 62, pp. 148–158, 2016.

[6] D. Mahata, J. R. Talburt, and V. K. Singh, "From chirps to whistles: Discovering event-specific informative content from twitter," in *Proceedings of the ACM Web Science Conference*. ACM, 2015, p. 17.

[7] A. Sarker and G. Gonzalez-Hernandez, "Overview of the second social media mining for health (smm4h) shared tasks at amia 2017," *Training*, vol. 1, no. 10,822, p. 1239.

[8] D. M. Jasper Friedrichs and S. Gupta, "Infynlp at smm4h task 2: Stacked ensemble of shallow convolutional neural networks for identifying personal medication intake from twitter," *Proceedings of the Second Workshop on Social Media Mining for Health Applications (SMM4H). Health Language Processing Laboratory.*, 2017.

[9] J. A. Cramer, R. H. Mattson, M. L. Prevey, R. D. Scheyer, and V. L. Ouellette, "How often is medication taken as prescribed?: A novel assessment technique," *Jama*, vol. 261, no. 22, pp. 3273–3277, 1989.

[10] D. Saparova, "Motivating, influencing, and persuading patients through personal health records: a scoping review," *Perspectives in Health Information Management/AHIMA, American Health Information Management Association*, vol. 9, no. Summer, 2012.

[11] E. Cambria, T. Benson, C. Eckl, and A. Hussain, "Sentic proms: Application of sentic computing to the development of a novel unified framework for measuring health-care quality," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 533–10 543, 2012.

[12] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[13] Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu, "Deep learning for healthcare decision making with emrs," in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 556–559.

[14] B. Shin, T. Lee, and J. D. Choi, "Lexicon integrated cnn models with attention for sentiment analysis," *arXiv preprint arXiv:1610.06272*, 2016.

[15] M. Denkowski and G. Neubig, "Stronger baselines for trustable results in neural machine translation," *arXiv preprint arXiv:1706.09733*, 2017.

[16] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[17] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[18] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, and L. Feng, "User-level psychological stress detection from social media using deep neural network," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 507–516.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[20] H. T. Le, C. Cerisara, and A. Denis, "Do convolutional networks need to be deep for text classification?" *arXiv preprint arXiv:1707.04108*, 2017.

[21] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. ACM, 2010, pp. 37–44.

[22] F. J. Grajales III, S. Sheps, K. Ho, H. Novak-Lauscher, and G. Eysenbach, "Social media: a review and tutorial of applications in medicine and health care," *Journal of medical Internet research*, vol. 16, no. 2, 2014.

[23] A. Sarker, K. OConnor, R. Ginn, M. Scotch, K. Smith, D. Malone, and G. Gonzalez, "Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from twitter," *Drug safety*, vol. 39, no. 3, pp. 231–240, 2016.

[24] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 502–518.

[25] A. Sarker, R. Ginn, A. Nikfarjam, K. OConnor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez, "Utilizing social media data for pharmacovigilance: A review," *Journal of biomedical informatics*, vol. 54, pp. 202–212, 2015.

[26] R. M. Bell, Y. Koren, and C. Volinsky, "All together now: A perspective on the netflix prize," *Chance*, vol. 23, no. 1, pp. 24–29, 2010.

[27] H. Wang, T. Zhao, H. Tan, and S. Zhang, "Biomedical named entity recognition based on classifiers ensemble," *IJCSA*, vol. 5, no. 2, pp. 1–11, 2008.

[28] A. Ekbal and S. Saha, "Stacked ensemble coupled with feature selection for biomedical entity extraction," *Knowledge-Based Systems*, vol. 46, pp. 22–32, 2013.

[29] U. K. Sikdar, A. Ekbal, and S. Saha, "Differential evolution based feature selection and classifier ensemble for named entity recognition," *Proceedings of COLING 2012*, pp. 2475–2490, 2012.

[30] R. Speck and A.-C. N. Ngomo, "Ensemble learning for named entity recognition," in *International semantic web conference*. Springer, 2014, pp. 519–534.

[31] K. Dinakar, E. Weinstein, H. Lieberman, and R. L. Selman, "Stacked generalization learning to analyze teenage distress," in *ICWSM*, 2014.

[32] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: many could be better than all," *Artificial intelligence*, vol. 137, no. 1-2, pp. 239–263, 2002.

[33] L. Deng and J. Platt, "Ensemble deep learning for speech recognition," pp. 1915–1919, 2014.

[34] J. Liu, S. Zhao, and X. Zhang, "An ensemble method for extracting adverse drug events from social media," *Artificial intelligence in medicine*, vol. 70, pp. 62–76, 2016.

[35] F. Godin, B. Vandersmissen, W. De Neve, and R. Van de Walle, "Multimedia lab@ acl w-nut ner shared task: named entity recognition for twitter microposts using distributed word representations," *ACL-IJCNLP*, vol. 2015, pp. 146–153, 2015.

[36] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.