

## Research Article

# Detecting Personal Medication Intake in Twitter via Domain Attention-Based RNN with Multi-Level Features

Shufeng Xiong <sup>1</sup>, Vishwash Batra,<sup>2</sup> Liangliang Liu,<sup>1</sup> Lei Xi <sup>1</sup> and Changxia Sun <sup>1</sup>

<sup>1</sup>Henan Agricultural University, Zhengzhou 450002, China

<sup>2</sup>School of Computing and Mathematics, Keele University, Keele ST55AA, UK

Correspondence should be addressed to Changxia Sun; [sunchangxia@henau.edu.cn](mailto:sunchangxia@henau.edu.cn)

Received 11 June 2022; Revised 8 July 2022; Accepted 13 July 2022; Published 9 August 2022

Academic Editor: Dong Chen

Copyright © 2022 Shufeng Xiong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Personal medication intake detection aims to automatically detect tweets that show clear evidence of personal medication consumption. It is a research topic that has attracted considerable attention to drug safety surveillance. This task is inevitably dependent on medical domain information, and the current main model for this task does not explicitly consider domain information. To tackle this problem, we propose a domain attention mechanism for recurrent neural networks, LSTMs, with a multi-level feature representation of Twitter data. Specifically, we utilize character-level CNN to capture morphological features at the word level. Subsequently, we feed them with word embeddings into a BiLSTM to get the hidden representation of a tweet. An attention mechanism is introduced over the hidden state of the BiLSTM to attend to special medical information. Finally, a classification is performed on the weighted hidden representation of tweets. Experiments over a publicly available benchmark dataset show that our model can exploit a domain attention mechanism to consider medical information to improve performance. For example, our approach achieves a precision score of 0.708, a recall score of 0.694, and a *F1* score of 0.697, which is significantly outperforming multiple strong and relevant baselines.

## 1. Introduction

Social media (Twitter, Facebook, etc.) encourages users to frequently express their thoughts, opinions, and other personal information of their lives. Existing studies have demonstrated that social media messages can provide knowledge for tracking public political opinion [1], detecting news events [2], and tracking the spread of infectious diseases [3]. Some research has also shown that social media can be used as a resource for mining public health information [4–6], especially in cases where the health data from official institutions is not readily available.

In this work, we focus on the task of medication intake detection, which is an individual level surveillance of drug safety [7] in the public health domain. The goal of the task is to automatically detect tweets that express clear evidence of personal medication consumption. For example, messages like *indicate the author intakes medicine while indicates no intake.*

There are two types of existing methods for this emerging task in the public health domain. The first involves applying traditional classification algorithms with hand-crafted features. For example, Kiritchenko et al. [8] exploited an SVM classifier with a variety of features for this task. Another one excludes the time-consuming effort of designing hand-crafted features by using deep neural networks and directly feeding word embedding as the input. Several methods based on CNNs (Convolutional Neural Networks) have achieved good performance [9–11].

Both of them have limitations. Traditional classification methods lack nonlinear mapping ability, although they can make full use of domain knowledge information (e.g., drug lexicon and domain word clusters) [8]. The neural-based methods based on pre-trained embeddings of words [12, 13] are domain-independent and are not very effective for a specific task [14, 15], for example, our medicine intake classification task. In addition, it is not always possible to train task-specific word embeddings due to limitations on

training resources. Despite some efforts to solve the problem of domain relevance at the feature level, for example, sentiment analysis on Nepali COVID-19 tweets [16], no relevant research work considering domain information has been found on our task.

To deal with the aforementioned limitations, we introduce a domain attention mechanism for recurrent neural networks with multi-level inputs to learn an informative representation of tweets. The attention mechanism enables the model the ability of learning domain-specific (medicine) matrix representation, which automatically weights the words in the text accordingly in the medication intake detection task. Meanwhile, the proposed model considers both word-level and character-level features as input features of the network. A prominent advantage of using character-level representation is that it is beneficial for many text analysis tasks [17–19], especially for informal text [20, 21], for example, tweets.

In particular, the proposed model generates word representations using a character-level CNN, which are fed to a highway network. We then concatenate them with pre-trained word embeddings, before feeding them to a BiLSTM network. Subsequently, as previously mentioned, the BiLSTM is introduced with an attention mechanism to distinctively attend on different words while learning the representation of the text. The attention-based BiLSTM also learns the representation of higher-level features in the whole text sequence of a tweet. Finally, softmax is applied to the final tweet representation for the classification task. We compared the experimental results obtained using our method with several strong and relevant baselines. We observe that our approach, with a micro-averaged F-score of 0.697 for Classes 1 and 2, achieves better performance on all other methods except ensemble approaches, which are more efficient than the standard approach. Altogether, this work introduces a novel attentional RNN framework with multi-level features that can effectively be applied to the personal medication intake detection task.

## 2. Related Work

Personal medication intake detection belongs to a short text classification task. Traditional representative methods for this task include statistical machine learning methods and deep learning methods. The vast majority of the first category is based on the vector space model, which is a typical method for tweet classification [22, 23]. Wang et al. [24] developed an SVM-based text classification algorithm. Chen et al. [25] and Jiang et al. [26] exploited the Naive Bayesian (NB) approach and KNN for this task, respectively. Wan et al. [27] implemented a new document classification method by integrating KNN and SVM, while Rogati et al. [28] investigated a large number of feature selection methods for text classification. However, these methods heavily depend on feature engineering, which cannot represent the grammatical and deep semantic information of words well.

Deep learning methods can automatically select features and therefore have become the mainstream methods for text

classification in recent years. The first step is to learn word representations using related methods [29–31]. Based on them, researchers initially adopted the CNN-based method to classify texts [32, 33]. Collobert et al. [33] extracted local features by using a convolutional layer. Kim [34] constructed a single-layer convolution network for sentence classification. Kalchbrenner et al. [35] proposed a CNN model with multi-layer dynamic k-Max pooling, taking random low-dimensional word vectors as input. Er et al. [36] developed an attention-based pooling component, which has the ability to obtain more semantic information. Yin et al. [37] developed a multi-channel variable-size CNN, which can support multiple pre-trained word embeddings and variable-size convolution kernel to obtain multi-granularity phrase features. Recently, the RNN-based model shows good performance. Lee et al. [38] exploited a convolutional recurrent neural network to process long text sequences. Lai et al. [39] proposed a bi-directional recurrent structure that can utilize the context information of words to classify text.

In addition, the participating systems of the SMM4H shared task are related to our method. These systems can be also divided into traditional statistic methods [8, 40] and neural network methods [10, 41]. Due to the characteristics of pursuing high-performance scores in evaluation tasks, most of them used ensemble technology. More details can be referred to the literature [9].

## 3. Background

*3.1. Personal Medication Intake Detection.* The primary objective of the personal medication intake detection task is the automatic classification of tweets mentioning medication intake, which is an emerging research topic in the public health domain based on social media. This is a three-class text classification task. Each medicine-mentioned tweet needs to be grouped into one of three categories: definite intake, non-intake, and possible intake. The details of these categories are as follows.

- (i) Define intake (Class 1)-The user expresses clear evidence of personal medication consumption, for example, “Benadryl and Tylenol are the only things saving me at night these last few nights.”
- (ii) Possible intake (Class 2)-It is suggested that one poster may have taken the medication, but there is no clear evidence, for example, “I would love to intravenously pump Motrin and caffeine into my body immediately.”
- (iii) Non-intake (Class 3)-There is no evidence showing that the user has consumed the medication, while it only mentions medication names, for example, “stay out of the heat, only drink water, and stay off your feet for a day or two. Tylenol is all you can take for pain.”

*3.2. Character Convolutional Neural Network (CNN).* Character Convolutional Neural Network (C-CNN) [17, 42] is fed characters instead of words, as in traditional CNN.

Given an input word  $w$ , which can be seen as a character sequence  $C = \{c_1, c_2, \dots, c_n\}$ , where  $n$  is the length of the word. The C-CNN applies the convolution operation on the character sequence to generate the feature map  $V_1$  as follows:

$$V_1 = \text{Conv}(W, C) + b, \quad (1)$$

where Conv denotes the convolution kernel and  $W$  and  $b$  are learnable parameters. In practice, there are different convolution kernels for catching various features. A pooling layer, which is utilized to compress and obtain crucial features for the next layer, is usually applied after the convolution layer. The computing process can be written as

$$V_2 = \text{Pooling}(V_1). \quad (2)$$

There are two common pooling operations: max pooling and mean pooling. For example, max pooling chooses the maximum value in a pooling window as the output result of the pooling process. Several combinations of convolution and pooling layers could be used in practice for specific tasks.

### 3.3. Bi-Directional Long Short Term Memory (BiLSTM).

The Long Short Term Memory (LSTM) network was introduced by Hochreiter et al. [43] and was refined and promoted by many works [44–46]. The LSTM solves the long-term dependency problem in the RNN model [47]. Given a sequence  $X = \{x_1, x_2, \dots, x_n\}$  as input, the operations performed by the LSTM units are as follows:

$$\begin{aligned} s_t &= [h_{t-1}, x_t], \\ f_t &= W_f \cdot s_t + b_f, \\ i_t &= W_i \cdot s_t + b_i, \\ u_t &= W_u \cdot s_t + b_u, \\ C_t &= \delta(f_t) * C_{t-1} + \delta(i_t) * \tanh(u_t), \\ o_t &= W_o \cdot (s_t + b_o), \\ h_t &= \delta(o_t) * \tanh(C_t), \end{aligned} \quad (3)$$

where  $h_t$  is the output of the LSTM at time step  $t$ .  $W$  and  $b$  are the weights and bias, respectively, and  $\delta$  is a sigmoid layer. In many NLP tasks, a bi-directional LSTM is used to obtain forward and backward information of words in a sequence. In BiLSTMs, it concatenates the outputs of the forward and backward hidden states as its output:

$$h_t = \left[ \vec{h}_t, \overleftarrow{h}_t \right]. \quad (4)$$

## 4. Methods

Our proposed model combines the C-CNN and BiLSTM. It also introduces an attention mechanism in the BiLSTM for the personal medication intake detection task. The model consists of a character-level word embedding component, a word-level feature representation component (Character Language Model, CLM) that uses C-CNN, a sentence-level

feature representation component using BiLSTM, and a Domain Attention Component (DAC). An overview of our model is shown in Figure 1.

Since tweets are mostly informal, traditional word embeddings cannot represent it well. Therefore, we use a Character Language Model (CLM) to capture morphological features at the word level. Firstly, a character embedding  $e_c$  is created for each character in a word. Our model then converts the character embedding sequence into a vector using a CLM, which is a kind of C-CNN network. The structure of the CLM is as shown on the right in Figure 1.

Specifically, for every word  $w$  in a sentence, after passing it to convolutional and max pooling layers, our model utilizes a highway network[42, 48] to regulate the information flow:

$$\begin{aligned} \widehat{C} &= g(W_h \cdot V_2 + b_h) \cdot t + V_2 \cdot (1 - t), \\ t &= H(W_t \cdot V_2 + b_t), \end{aligned} \quad (5)$$

where  $H$  is a nonlinear function,  $V_2$  is calculated by equation (2), and  $t$  and  $(1 - t)$  are called the transform gate and carry gate, respectively.

After obtaining  $\widehat{C}_i$ , the representation of the  $i$ -th word at character-level, we concatenate it with its word embedding  $e_i$  to generate the final representation of the word:

$$\widehat{e}_i = [\widehat{C}_i, e_i]. \quad (6)$$

Subsequently, we feed a sentence  $s = (\widehat{e}_1, \widehat{e}_2, \widehat{e}_3, \dots, \widehat{e}_n)$  into a BiLSTM network to get the hidden states  $h = (h_1, h_2, h_3, \dots, h_n)$ . In our experiments, we treat each tweet as one sentence and yet achieve good results since most of the tweets in our dataset are too short and mostly contain one or two sentences.

At this stage, the model performs a general processing on tweets. Therefore, for the medicine intake detection task, we introduce a DAC to attend to the specific domain information that is being used to detect a specific condition. The DAC aims to weigh the informative words for medicine intake highly. First, the result  $\widehat{h}_k$  obtained by inputting  $h_k$  into a single-layer perceptron is used as the hidden representation of  $h$ . The weight value of a word is determined by the similarity of  $\widehat{h}_k$  and a parameter  $D$ , here  $D$  can be seen as a domain context vector. After processing using a softmax function, a normalized attention weight matrix is obtained, which indicates the weight of each word in a sentence. Finally, the tweet representation  $u$  can be calculated as the weighted summation of the words in it. The output is computed as follows:

$$\begin{aligned} \widehat{h}_k &= \tanh(W_a \cdot (h_k + b_a)), \\ a_k &= \frac{\exp(\widehat{h}_k^\top \cdot D)}{\sum_k (\widehat{h}_k^\top \cdot D)}, \\ u &= \sum_k a_k \cdot \widehat{h}_k, \end{aligned} \quad (7)$$

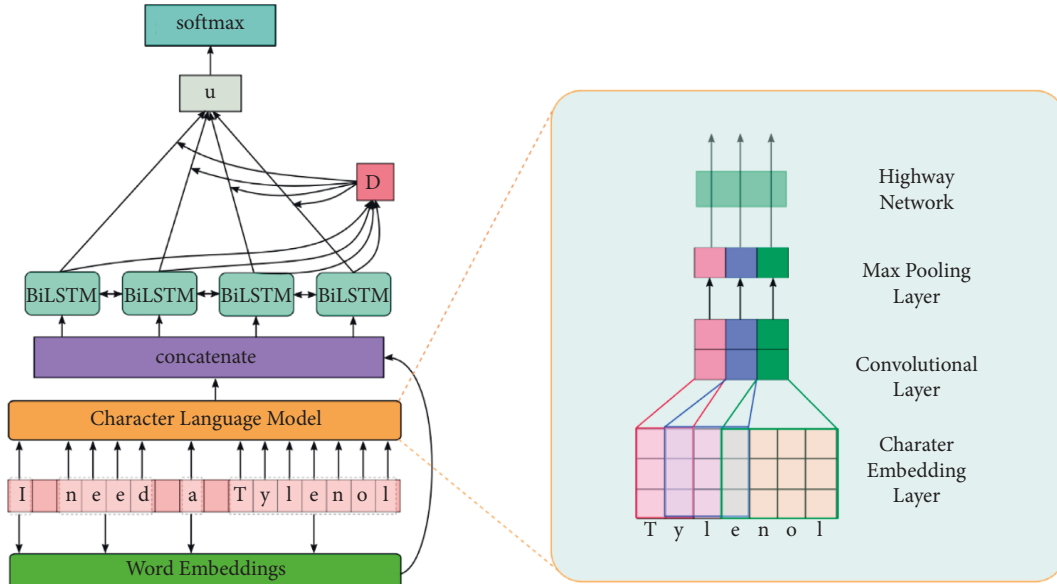


FIGURE 1: Framework of our model.

where  $W_a$  and  $b_a$  are the weight and bias, respectively.  $a_k$  stands for the attention value of the  $k$ -th word and measures the weight of each word in the sentence.

The vector representing the whole text sequence from a tweet or the tweet vector  $u$  is a higher-level representation and can be used directly as a feature for medicine intake detection:

$$p = \text{softmax}(W_s \cdot (u + b_s)). \quad (8)$$

The final optimization objective is to minimize the negative log likelihood of the correct labels:

$$L = - \sum_d \hat{p} \log p, \quad (9)$$

where  $\hat{p}$  represents the ground-truth label of the tweet.

## 5. Experiments

**5.1. Dataset.** Our experiments were conducted on a publicly available dataset from the 2nd SMM4H (Social Media Mining for Health) Shared Task on the AMMIA 2017 website. Using the Twitter download script and the tweet dataset description file provided by the organizers, we did not collect all the tweets since some of them are not available. Table 1 summarizes the statistics for the dataset. Classes 1, 2, and 3 stand for personal medication intake, possible medication intake, and no medication intake, respectively. Our training dataset is a combination of the originally provided training and validation datasets. We utilize 10-fold cross-validation when training our model.

**5.2. Model Configuration and Training.** We use pre-trained word embeddings to initialize the input of the neural network model. This is highly useful for NLP tasks [49, 50]. In our work, we use the embeddings trained by a word2vec

TABLE 1: Statistics of the dataset.

	Class 1	Class 2	Class 3	Total
Train	1648	2650	4246	8544
Test	1444	2267	2593	6304

model on Twitter data [51], which are of 400 dimensions. For character embeddings, we use a random initialization since there are no publicly available character embeddings in this case.

Within our experiments, we have two types of parameters, hyper-parameters and other settings. Specifically, the character embedding dimension is 15, the dimension of the hidden layer is 300, and the CLM has filters of width [1, 2, 3, 4] of size [15, 30, 45, 60] for a total of 180 filters. Additionally, the batch size, the learning rate, the dropout rate, and the  $L_2$  normalization factor are set to 100,  $3e-4$ , 0.3, and  $5e-7$ , respectively. In our training process, we used early stopping with a patience value of 40.

**5.3. Baselines.** We conducted comparative studies involving experiments with several baseline methods on the dataset, including neural network methods, traditional machine learning algorithms, and state-of-the-art methods for this task. In the first category, we choose the NB and SVM algorithms:

NB is a Naive Bayes classifier in which  $n$ -grams ( $n = 1, 2, 3$ ) are used as features.

SVM is a Support Vector Machine classifier with  $n$ -grams ( $n = 1, 2, 3$ ) features.

The neural network model is currently the dominant method for text processing. We chose the following representative methods:

BiLSTM uses a traditional bi-directional LSTM model for medicine intake detection, which represents a sentence with the hidden state of the last word of it.

CharCNN [18] is a classical model which performs text classification by using a character-level convolutional network.

AttRNN [52] concatenates the last hidden state, the first hidden states of an RNN with an attentive representation of a hidden state sequence as the features of a text sequence.

The third group is the top three systems from the SMM4H Shared Task:

InfyNLP [10] is the first system in the 2nd SMM4H Shared Task at AMIA 2017. It uses a stacked ensemble of shallow CNNs modeled as a classifier for this task.

UKNLP [41] is the second system in the Shared Task which utilizes a CNN network with a self-attention component.

NRC-Canada [8] exploited the SVM classifier with a variety of hand-crafted features, which is the third system on the SMM4H Shared Task.

**5.4. Results and Discussion.** Table 2 presents the performance of the different methods. We presented the micro-averaged Precision, Recall, and *F1* scores of Class 1 (personal medication intake) and Class 2 (possible medication intake). The best results are shown in bold text. The results with “#” are copied from their original papers. Following the setting in the SMM4H shared task, we report the micro-averaged scores over Class 1 and Class 2. It was observed that the proposed model performs the best over the *F1* score against strong baselines and top systems in the SMM4H shared task. Compared with other neural classification methods, the proposed domain attention component and CLM in our model improve the performance in the context of the task at hand. CharCNN and AttRNN methods perform less than our method, while BiLSTM performs poorly in this group because they are proposed for general text classification tasks, for example, topic classification. The former two methods perform better than BiLSTM because they introduce character-level information and attention components. NB and SVM, as they are represented in traditional machine methods, perform poorly because they cannot fully capture text semantic information compared to the NN model. At the performance of the top three systems is not as good as our method.

**5.5. Ablation Test.** In this subsection, we discuss the impact and contribution of the different components of our model. Specifically, we tested 3 settings. The first, we dismissed CLM only. In this case, the model did not capture character-level features. In the next setting, we remove the DAC only. Similarly, the model did not care about the domain information. Finally, we dismiss both CLM and DAC. In this case, the model degenerates to BiLSTM, which only just uses word-level features via BiLSTM encoding. Table 3 reports the results of this ablation study.

TABLE 2: Comparison results.

Model	Precision	Recall	<i>F1</i>
NB	0.675	0.631	0.650
SVM	0.679	0.664	0.668
BiLSTM	0.683	0.672	0.678
CharCNN	0.681	0.697	0.689
AttRNN	0.704	0.677	0.688
InfyNLP	0.725 <sup>#</sup>	0.664 <sup>#</sup>	0.693 <sup>#</sup>
UKNLP	0.701 <sup>#</sup>	0.677 <sup>#</sup>	0.689 <sup>#</sup>
NRC-Canada	0.704 <sup>#</sup>	0.635 <sup>#</sup>	0.668 <sup>#</sup>
Our model	0.708	0.694	0.697

TABLE 3: Comparison results with different settings.

Model	Precision	Recall	<i>F1</i>
w/o CLM and DAC	0.683	0.672	0.678
w/o CLM	0.701	0.683	0.693
w/o DAC	0.669	0.713	0.691
Full model	0.708	0.694	0.697

It is clear that both CLM and DAC are critical to the performance of our model. Removing one or both of them can cause performance degradation. In particular, we also observe that CLM seems to be less important than DAC, which means that the performance drops more as compared to removing CLM.

## 6. Conclusion and Future Works

Personal medication intake detection, aiming to automatically detect tweets that express clear evidence of personal medication consumption, is an essential research topic in the surveillance of drug safety. In this work, we proposed a domain attention component for recurrent neural networks, for example, LSTM, with multi-level feature representation of text from Twitter. Through experiments on the public benchmark dataset, we validated the performance of our model. Our model obtains the best performance of 0.697 *F1* score. Compared with multiple strong baselines, it showed a significant performance improvement.

Our method still has limitations on domain-specific knowledge representation due to the representation ability of the neural network model itself. Thus, it would be interesting to combine the knowledge base, for example, Knowledge Graph, with our model to obtain richer domain information for this task.

## Data Availability

The data used to support the findings of this study have been deposited in the website <https://healthlanguageprocessing.org/sharedtask2/smm4h-sharedtask-2017/>

## Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

## Acknowledgments

The author shall be thankful to Pingdingshan University for providing resources for this work. This work was supported in part by the MOE (Ministry of Education of China) Project of Humanities and Social Sciences (No. 19YJCZH198) and the Science and Technology Planning Project of Henan Province, China (No. 222102110423).

## References

- [1] B. O'Connor, R. Balasubramanian, B. R. Routledge, and N. A. Smith, "From tweets to polls: linking text sentiment to public opinion time series," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Menlo Park, AAAI 2010.
- [2] A.-M. Popescu and M. Pennacchiotti, "Detecting controversial events from twitter," in *Proceedings of the 19th ACM international conference on Information and knowledge management. ACM*, pp. 1873–1876, Ontario, Canada, October 2010.
- [3] A. Lamb, M. J. Paul, and M. Dredze, "Separating fact from fear: tracking flu infections on twitter," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 789–795, Stroudsburg, PA, USA, 2013.
- [4] L. Liu, W. Kunyan, Z. Xingting, D. Weng, L. Gao, and J. Lei, "The current status and a new approach for Chinese doctors to obtain medical knowledge using social media: a study of WeChat," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 2329876, , 2018.
- [5] L. Sinnenberg, A. M. Buttenheim, K. Padrez, C. Mancheno, L. Ungar, and R. M. Merchant, "Twitter as a tool for health research: a systematic review," *American Journal of Public Health*, vol. 107, no. 1, pp. 143–e8, 2017.
- [6] C. C. Freifeld, J. S. Brownstein, C. M. Menone et al., "Digital drug safety surveillance: monitoring pharmaceutical products in twitter," *Drug Safety*, vol. 37, no. 5, pp. 343–350, 2014.
- [7] 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," in *BioNLP 2017*, pp. 136–142, Association for Computational Linguistics, Stroudsburg, PA, USA, 2017.
- [8] S. Kiritchenko, S. M. Mohammad, J. Morin, and B. de Bruijn, "NRC-Canada at SMM4H shared task: classifying tweets mentioning adverse drug reactions and medication intake," in *Proceedings of the Social Media Mining for Health Applications Workshop at AMIA-2017*, Washington DC, DC, December 2017.
- [9] A. Sarker, M. Belousov, J. Friedrichs et al., "Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1274–1283, 2018.
- [10] J. Friedrichs, D. Mahata, and S. Gupta, "InfyNLP at SMM4H task 2:stacked ensemble of shallow convolutional neural networks for identifying personal medication intake from twitter," in *Proceedings of the Second Workshop on Social Media Mining for Health Applications (SMM4H)*, Washington, DC, USA, 2017.
- [11] D. Mahata, J. Friedrichs, R. R. Shah, and J. Jiang, "Detecting personal intake of medicine from twitter," *IEEE Intelligent Systems*, vol. 33, no. 4, pp. 87–95, 2018.
- [12] H. Schwenk, "Continuous space language models," *Computer Speech & Language*, vol. 21, no. 3, pp. 492–518, 2007.
- [13] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [14] P. Fu, Z. Lin, F. Yuan, W. Wang, and D. Meng, "Learning sentiment-specific word embedding via global sentiment representation," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pp. 4808–4815, Louisiana, LA, USA, February 2018.
- [15] H. Zamani and W. B. Croft, "Relevance-based word embedding," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 505–514, New York, NY, USA, August 2017.
- [16] Y. Ren, R. Wang, and D. Ji, "A topic-enhanced word embedding for Twitter sentiment classification," *Information Sciences*, vol. 369, pp. 188–198, nov 2016.
- [17] C. N. Dos Santos and B. Zadrozny, "Learning character-level representations for part-of-speech tagging," in *Proceedings of the 31st International Conference on International Conference on Machine Learning*, Beijing, China, June 2014.
- [18] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional net-works for text classification," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., pp. 649–657, Curran Associates, Inc, New York, NY, USA, 2015.
- [19] R. Kavuluru, A. Rios, and T. Tran, "Extracting drug-drug interactions with word and character-level recurrent neural networks," in *Proceedings of the 2017 IEEE International Conference on Healthcare Informatics, ICHI*, Park City, Utah, USA, August 2017.
- [20] S. Vosoughi, P. Vijayaraghavan, and D. Roy, "Tweet2Vec: learning tweet embeddings using character-level CNN-LSTM encoder-decoder," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1041–1044, Pisa, Italy, July 2016.
- [21] D. Liang, W. Xu, and Y. Zhao, "Combining word-level and character-level representations for relation classification of informal text," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 43–47, Vancouver, Canada, August 2017.
- [22] C. Sitaula, A. Basnet, A. Mainali, and T. B. Shahi, "Deep learning-based methods for sentiment analysis on Nepali COVID-19-related tweets," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 2158184, 2021.
- [23] T. B. Shahi, C. Sitaula, and N. Paudel, "A hybrid feature extraction method for Nepali COVID-19-related tweets classification," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–11, Article ID 5681574, 2022.
- [24] Z. Q. Wang, X. Sun, D. X. Zhang, and X. Li, "An optimal SVM-based text classification algorithm," in *Proceedings of the 2006 International Conference on Machine Learning and Cybernetics*, pp. 1378–1381, Dalian, China, August 2006.
- [25] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," *Expert Systems with Applications*, vol. 36, pp. 5432–5435, 2009.
- [26] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved K-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1503–1509, 2012.
- [27] C. H. Wan, L. H. Lee, R. Rajkumar, and D. Isa, "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine,"

- Expert Systems with Applications*, vol. 39, no. 15, Article ID 11880, 2012.
- [28] M. Rogati and Y. Yang, "High-performing feature selection for text classification," in *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pp. 659–661, Virginia VA, USA, November 2002.
- [29] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.
- [30] J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proceedings of the EMNLP 2014*, pp. 1532–1543, Doha, October 2014.
- [31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the NIPS*, pp. 3111–3119, Nevada, NV, USA, December 2013.
- [32] C. Sitaula and T. B. Shahi, "Multi-channel CNN to classify Nepali covid-19 related tweets," 2021, <https://arxiv.org/>.
- [33] R. Collobert, J. Weston, L. E. O. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.
- [34] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, October 2014.
- [35] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the ACL*, Maryland, MD, USA, June 2014.
- [36] M. J. Er, Y. Zhang, N. Wang, and M. Pratama, "Attention pooling-based convolutional neural network for sentence modelling," *Information Sciences*, vol. 373, pp. 388–403, 2016.
- [37] W. Yin and H. Schütze, "Multichannel variable-size convolution for sentence classification," in *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pp. 204–214, Beijing, China, July 2015.
- [38] J. Y. Lee and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 515–520, San Diego, California, CA, USA, 2016.
- [39] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proceedings of the Twenty-ninth AAAI conference on artificial intelligence*, Texas, TX, USA, January 2015.
- [40] B. G. Hb and S. Kp, "NLP\_CEN\_AMRITA @ SMM4H:health care text classification through class embeddings," in *Proceedings of the Social Media Mining for Health Applications Workshop at AMIA-2017*, Washington, DC, USA, 2017.
- [41] S. Han, T. Tran, A. Rios, and R. Kavuluru, "Team UKNLP: detecting ADRs, classifying medication intake messages, and normalizing ADR mentions on twitter," in *Proceedings of the CEUR Workshop*, pp. 49–53, Switzerland, August 1996.
- [42] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Arizona, AZ, USA, March 2016.
- [43] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [44] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional LSTM net-works for improved phoneme classification and recognition," in *Proceedings of the ICANN 2005*, pp. 799–804, Warsaw, Poland, September 2005.
- [45] F. A. Gers, J. Schmidhuber, and F. A. Cummins, "Learning to forget: continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
- [46] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Proceedings of the INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, pp. 194–197, Portland, Oregon, OR, USA, September 2012.
- [47] B. A. Pearlmutter, "Learning state space trajectories in recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 263–269, 1989.
- [48] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., pp. 2377–2385, Curran Associates, Inc, New York, NY, USA, 2015.
- [49] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan, and G. Neubig, "When and why are pre-trained word embeddings useful for neural machine translation?" in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 529–535, New Orleans, Louisiana, 2018.
- [50] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Double Embeddings and CNN-Based Sequence Labeling for Aspect Extraction," in *in Proceedings of the ACL*, vol. 2, pp. 592–598, Melbourne, Australia, 2018.
- [51] F. Godin, B. Vandersmissen, W. De Neve, and R. De Walle, "MultimediaLab @ ACL WNUT NER shared task: named entity recognition for twitter microposts using distributed word representations," in *Proceedings of the Workshop on Noisy User-generated Text*, pp. 146–153, Beijing, China, July 2015.
- [52] C. Du and L. Huang, "Text classification research with attention-based recurrent neural networks," *International Journal of Computers, Communications & Control*, vol. 13, no. 1, pp. 50–61, 2018.