ULD-NUIG at Social Media Mining for Health Applications (#SMM4H) Shared Task 2021

Atul Kr. Ojha, Priya Rani, Koustava Goswami,
Bharathi Raja Chakravarthi, John P. McCrae
Data Science Institute, National University of Ireland Galway
(atulkumar.ojha, priya.rani, koustava.goswami,
bharathi.raja, john.mccrae)@insight-centre.org

Abstract

In this paper, we present the ULD-NUIG team's system, designed as part of Social Media Mining for Health Applications (#SMM4H) Shared Task 2021. We participate in two tasks out of eight, namely "Classification of tweets self-reporting potential cases of COVID-19" (Task 5) and "Classification of COVID19 tweets containing symptoms" (Task 6). The team conduct a series of experiments to explore the challenges of both the tasks. We used a multilingual pre-trained BERT model for Task 5 and Generative Morphemes with Attention (GenMA) model for Task 6. In the experiments, we find that, GenMA, developed for Task 6, gives better results on both validation and test data-set. The submitted systems achieve F-1 score 0.53 for Task 5 and 0.84 for Task 6 on test data-set.

1 Introduction

In recent decades, social media has proved to be one of the greatest sources of information exchange. When the world was overtaken by the COVID-19 outbreak, social media became the greatest platform for the general public to exchange different information about the pandemic. With the widespread digitization of behavioural and medical data, the emergence of social media, and the Internet's infrastructure of large-scale knowledge storage and distribution, there has been a breakthrough in our ability to identify human social interactions, behavioural patterns, cognitive processes and their relationships with healthcare. At the same time, it has also induced a different level of challenges in the natural language processing field such as detection of medical jargons, named entity recognition, multi-word expressions. Furthermore, the informal nature of tweets and short length, which often contain non-standard grammar, frequent misspellings, many contractions, extensive slang, and use of emojis/emoticons to express emotion exacerbate the challenges (Dang et al., 2020). The scenario becomes even more complicated since social media covers very large populations and the geographical location.

Despite several issues, social media data has been used to monitor human health and disease (the recent pandemic outburst) all over the world. Many promising methodologies are being developed. In this paper, we describe two different systems trained on the data provided by the Social Media Mining for Health Applications (#SMM4H) Shared Task 2021 organisers (Magge et al., 2021) namely Task 5: Classification of tweets self-reporting potential cases of COVID-19 and Task 6: Classification of COVID19 tweets containing symptoms (see Section 2). We conduct a series of experiments to explore the challenges of both the tasks. We use multilingual pre-trained BERT model for Task 5 and Generative Morphemes with Attention (GenMA) model for Task 6 (see Section 3).

2 Data Augmentation

2.1 Data Size

Task 5 : Classification of tweets self-reporting potential cases of COVID-19
We use the data set provided by the organisers of Task 5 SMM4H'21.[1] The data was divided into training set, validation set and test set as detailed in Table 1. The task involves binary classification of the tweets, which distinguishes self-reported potential cases of COVID-19 annotated as 1 and non potential cases annotated as 0.

Task 6 : Classification of COVID19 tweets containing symptoms
The data set for the experiment was given by the organisers of Task 6 SMM4H'21.[2] Like Task 5 this data was also divided in training,

[1] https://healthlanguageprocessing.org/smm4h-2021/task-5/
[2] https://healthlanguageprocessing.org/smm4h-2021/task-6/

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 149–152

validation and test set. The statistics of the data set is given in Table 1. The data is classified at three different levels: self-reports, non-personal reports, literature/news mentions.

Task	Training	Validation	Test
Task 5	6,465	717	10,000
Task 6	9,068	501	6,500

Table 1: Statistics of Task 5 and 6 Dataset

2.2 Pre-processing

We normalized the data through the following pre-processing steps as part of the experiment.

1. After a thorough manual evaluation of the data set, we came to the conclusion that emoticons, URLs along with the other special characters which are very common in social media data do not serve necessary purpose for our tasks. Therefore we removed emoticon, URLs and other special characters from the data set.

2. Lower casing all the tweets in the data set. After lower casing the tweets all extra spaces were removed from it.

3 Experiments

3.1 Task 5

We have used the multilingual pre-trained BERT (Devlin et al., 2019; Turc et al., 2019) model to fine-tune our model on the given Task 5 training data set. The detailed model descriptions is given below:

- The model has an embedding dimension of 768. We have used the Google-provided cased vocabulary.

- Parameters for training - We have trained our model for 3 epochs on the training data set and have used a stepped LR scheduler for the learning rate scheduling. The learning rate is set to 2e-5 (see Equation 1).

Based on Hugging Face implementation, we have used the below equation as warmup steps definition for training the model. Here 'r' is the tuneable parameter, which defines the percentage of data used to define the step size while training. We have used 10% of the data while training. After training the model we have tested it on the held-out test data set given by the organizers.

$$W_{steps} = \frac{(len(training_{set}) * epochs_{training})}{batchsize_{training} * r} \quad (1)$$

3.2 Task 6

We have taken the inspiration from the Generative Morphemes with Attention (GenMA) model (Goswami et al., 2020) to develop the model for Task 6. We have noted the model description below:

- The model takes the character sequence as the input sequence. It has one character embedding layer and two convolutions (CONV1D) layers. Each convolution layer has one max-pooling layer. After the convolution layers, there is one LSTM layer and one bidirectional LSTM layer, followed by two self-attention layers. The model has two hidden layers and one softmax layer. The model generates new artificial morphemes and frames a sentence as a group of new morphemes. The combination of two CNN layers helps to generate new morphemes based on deep relative co-occurring characters (3 characters frame), and the LSTM layers help to capture the global information of sentences based on newly generated features. The self-attention layers help to construct sentence-level information. It also captures relativity strength among different co-occurring character features.

- We have used 32 filters, each with a kernel size of 3. The max-pooling size is 3. The hidden size h_i of LSTM units is kept to 100. The dense layer has 32 neurons, and it has 50 percent dropout. The Adam optimizer (Kingma and Ba, 2015) is used to train our model with the default learning set to 0.0001. The batch size is set to 10. For the convolution layer in both the experiments we have used the relu activation function (Nair and Hinton, 2010) and for the dense layer we have used tanh activation function (Kalman and Kwasny, 1992). Categorical cross-entropy loss is used for the multi-class classification. We have used Keras[3] to train and test our model.

[3] https://keras.io

The convolution layers act as the feature extractor of the sentences. The one-dimensional convolution implements one-dimensional filters which slide over the sentences as a feature extractor. The second convolution layer will take feature representations as input and generate a high-order feature representation of the characters. The max-pooling network after each convolution network helps to capture the most important features of size *d*. The new high-order representations are then fed to the LSTM (Long Short Term Memory Network) as input.

The LSTM layer takes the output of the previous CNN layer as input. The LSTM layer produces a new representation sequences in the form of $h_1, h_2, ...h_n$ where h_t is the hidden state of the LSTM of time step *t*, summarising all the information of the input features (morphemes) of the sentences. At each time step, *t*, the hidden state takes the previous time step hidden state h_{t-1} and characters (x_t) as input. A bidirectional LSTM (BiLSTM) network has been used, which has helped us to summarise the information of the features from both directions. The Bidirectional LSTM consists of a forward pass and a backward pass which provides two annotations of the hidden state h_{for} and h_{back}. We obtained the final hidden state representation by concatenating both hidden states $h_i = h_{i-for} \oplus h_{i-back}$, where h_i is the hidden state of the *i*-th timestep and $\oplus$ is the element-wise sum between the matrices.

The attention layer helps to determine the importance of one morpheme over others while building sentence embedding for classification. The self-attention mechanism has been adopted from Baziotis et al. (2018), which helped to identify the morphemes that capture the important features to classify the tweets. The mechanism assigns weight a_i to each feature's annotation based on output h_i of the BiLSTM's hidden states, with the help of the softmax function as illustrated in Equation 2 and 3 (Baziotis et al., 2018).

$$a_i = \tanh(W_h \cdot h_i + b_h) \tag{2}$$

$$a_i = \frac{exp(a_i)}{\sum_{t=1}^{T} exp(a_t))} \tag{3}$$

The new representation will give a fixed representation of the sentence by taking the weighted sum of all feature-label annotations as shown in Equation 4.

$$r = \sum_{i=1}^{T} a_i \cdot h_i \tag{4}$$

where W_h and b_h are the attention weights and bias respectively (Baziotis et al., 2018; Goswami et al., 2020).

The output layer consists of one fully-connected layer with one softmax layer. The sentence representation after the attention layer is the input for the dense layer. The output of the dense layer is the input of the softmax which gives the probability distribution of all the classes with the help of the softmax function.

4 Evaluation

We use shared task organizers' validation and test data-set for evaluation. The standard evaluation metrics, Precision, Recall and F-1 score, were used for automatic evaluation. It gives a quantitative picture of particular differences across different systems, especially with reference to evaluation scores. On the validation data-set, Task 5 and 6 systems' F-1 score were 0.89 and 0.95 respectively. While on the test data-set, F-1 score were 0.53 and 0.84 respectively. The detailed results are given in Table 2.

System	Precision	Recall	F-1 score
Task 5	0.7412	0.4091	0.53
Task 6	0.8415	0.8415	0.84

Table 2: Accuracy of Task-5 and 6 Systems on Test Data-set

5 Summing up

The entire series of experiments gave us various types of insights to deal with social media data for mining medical information. We observed that pre-trained language models such as BERT do not provide good results for extraction of medical information for COVID-19. One of the reasons that we could think is that these models are trained on various domain data set but it is very unlikely that these data-sets contain information regarding COVID-19. On the other hand our characters based attention model outperform the BERT model. In future, we would like to explore more models with word features, specific linguistic features in order to deeply understand the characteristics of social media mining for clinical information.

Acknowledgements

This publication has emanated from research in part supported by the EU H2020 programme under grant agreements 731015 (ELEXIS-European Lexical Infrastructure) as well as by the Irish Research Council under grant number SFI/18/CRT/6223 (CRT-Centre for Research Training in Artificial Intelligence) co-funded by the European Regional Development Fund.
We are also grateful to the organizers of SMM4H'21 Shared Task for providing us the Task 5 and 6 data and evaluation scores.

References

Christos Baziotis, Athanasiou Nikolaos, Pinelopi Papalampidi, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive RNNs. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 613–621, New Orleans, Louisiana. Association for Computational Linguistics.

Huong Dang, Kahyun Lee, Sam Henry, and Özlem Uzuner. 2020. Ensemble BERT for classifying medication-mentioning tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 37–41, Barcelona, Spain (Online). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Koustava Goswami, Priya Rani, Bharathi Raja Chakravarthi, Theodorus Fransen, and John P. McCrae. 2020. ULD@NUIG at SemEval-2020 task 9: Generative morphemes with an attention model for sentiment analysis in code-mixed text. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 968–974, Barcelona (online). International Committee for Computational Linguistics.

Barry L Kalman and Stan C Kwasny. 1992. Why tanh: choosing a sigmoidal function. In *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, volume 4, pages 578–581. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Arjun Magge, Ari Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulàlia Farré-Maduell, Salvador Lima López, Juan M Banda, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Martin Krallinger, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Overview of the Sixth Social Media Mining for Health Applications (# SMM4H) Shared Tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv e-prints*, pages arXiv–1908.