CSECU-DSG@SMM4H'22: Transformer based Unified Approach for Classification of Changes in Medication Treatments in Tweets and WebMD Reviews

Afrin Sultana[*], **Nihad Karim Chowdhury, and Abu Nowshed Chy**
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
afrin.sultana.cu@gmail.com, nihad@cu.ac.bd, and nowshed@cu.ac.bd

Abstract

Medications play a vital role in medical treatment as medication non-adherence reduces clinical benefit, results in morbidity, and medication wastage. Self-declared changes in drug treatment and their reasons are automatically extracted from tweets and user reviews, helping to determine the effectiveness of drugs and improve treatment care. SMM4H 2022 Task 3 introduced a shared task focusing on the identification of non-persistent patients from tweets and WebMD reviews. In this paper, we present our participation in this task. We propose a neural approach that integrates the strengths of the transformer model, the Long Short-Term Memory (LSTM) model, and the fully connected layer into a unified architecture. Experimental results demonstrate the competitive performance of our system on test data with 61% F1-score on task 3a and 86% F1-score on task 3b. Our proposed neural approach ranked first in task 3b.

1 Introduction

User-generated contents on social media and online health forums represent a wide variety of facts, experiences, and opinions on various health topics including personal health issues, reviews on medication, side effects, and informal questions on health concerns. Shared information on social media or online health forums allows to detect users' self-declared changes in medication. Self-declared changes include stopping a treatment, changing a dose, or forgetting to take the drugs, etc. In this work, we focus on classifying patients making changes to their medication treatments (i.e non-persistent patients) as a part of our participation in the Social Media Mining for Health (SMM4H) 2022 shared task 3 (Davy et al., 2022). This task includes two subtasks. We have to detect changes in medication treatments from tweets and WebMD reviews in subtasks 3a and 3b, respectively. WebMD is one of the top healthcare websites and an online publisher of news and information pertaining to drugs, human health and well-being. Table 1 represents some examples of persistent and non-persistent tweets and reviews from the SMM4H 2022 task 3 dataset, respectively.

Tweet/Review	Label
Task 3a: tweet classification	
E#1: I broke out with a rash after starting this medication, is this normal?	Persistent
E#2: This medication caused me to cough a lot . Hated it.	Non-persistent
Task 3b: WebMD review classification	
E#3: It's totally normal to take a ambien on a 2 hour flight, right?	Persistent
E#4: I quit Lipitor all together	Non-persistent

Table 1: Examples of persistent and non-persistent tweets/reviews.

The major contribution of this paper is that we propose an approach to explore the robustness of the RoBERTa (Liu et al., 2019) model with the unification of features from LSTM (Hochreiter and Schmidhuber, 1997) and a fully connected layer where RoBERTa maps the input text into meaningful embeddings effectively, LSTM captures long-term dependencies, and fully connected linear layer selects effective features to encapsulate context.

The rest of the paper describes our proposed framework, experimental details, and evaluation.

2 Proposed Framework

An overview of our proposed framework is shown in Figure 1. For a given text, we use the FLAIR (Akbik et al., 2019) framework to extract document embedding features from the transformer model RoBERTa. These features are carried out

[*] Corresponding Author

through two different capsules. One is a stacked LSTM and the other is a simple linear layer. Then, we concatenate the features of the LSTM and linear architecture, which are then applied to a feed-forward fully-connected output layer to procure predicted labels.

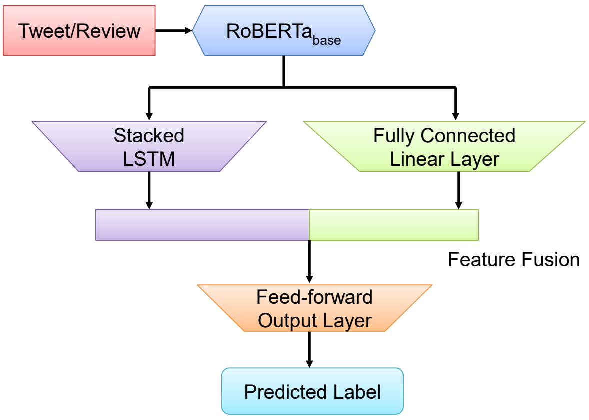


Figure 1: Proposed framework.

2.1 Transformer Document Embedding

Document Embedding provides an embed for the entire text. We leverage the FLAIR framework to generate document embeddings for each given text using a RoBERTa model from the Transformers (Vaswani et al., 2017) family.

RoBERTa: RoBERTa, a robustly optimized BERT pre-training method, is an extension to the original BERT (Devlin et al., 2018) model that trains the encoding vocabulary using larger byte-level bytes containing 50K subword units. It improves BERT by removing next-sentence prediction targets, dynamically changing mask patterns, and training the model with larger batches on more data (Liu et al., 2019).

2.2 Stacked LSTM

Document embedding vectors are fed into stacked long short-term memory (LSTM) capsules, as LSTMs capture long-term dependencies by storing previous information. Furthermore, stacked LSTM is an extension of the LSTM model that has multiple hidden LSTM layers, where each layer contains multiple memory cells. A stacked LSTM is utilized in order to increase capacity (Staudemeyer and Morris, 2019) and depth of the model (Graves et al., 2013). The stacked LSTM capsule generates an m-dimensional feature vector where ‘m’ is the hidden size of LSTM capsule.

2.3 Fully Connected Linear Layer

The document embedding vector obtained from RoBERTa is passed through a fully connected linear layer. Fully connected layers learn from high-level features and provide efficient feature representations that encapsulate the essence of a given high-level feature. An n-dimensional feature vector is produced from the document embedding vector where ‘n’ is the number of neurons of fully connected linear layer.

2.4 Output Layer

We concatenated the m-dimensional feature vector of the stacked LSTM capsule and the n-dimensional feature vector of the fully connected linear layer. Later, the fusion vector passes through a feed-forward fully connected linear layer operating as an output layer to obtain predicted labels. We utilize the SoftMax activation function in the output layer to normalize the features to probabilities, considering the highest probability class as the predicted label.

3 Experiments and Evaluations

3.1 Dataset Description and Evaluation Measures

Category	Train	Dev	Test
Task 3a: tweet classification			
Persistent	5380	1434	-
Non-persistent	518	138	-
Total	5898	1572	2360
Task 3b: WebMD review classification			
Persistent	4632	556	-
Non-persistent	5746	741	-
Total	10378	1297	1297

Table 2: The statistics of SMM4H 2022 task 3 dataset.

The organizers of the SMM4H 2022 task 3 (Davy et al., 2022) provided a benchmark dataset that consists of two corpora: a set of tweets with imbalanced positive and negative tweets, and a set of drug reviews from WebMD.com with balanced positive and negative reviews. Persistent data is labeled with ‘0’ whereas non-persistent data is labeled with ‘1’. Table 2 shows the statistics of the used dataset.

To evaluate the performance of participants’ systems, SMM4H 2022 task 3 organizers employed standard strategies for sub-tasks 3a and 3b. Standard evaluation metrics, including precision, recall, and F1-score for non-persistent classes, were used to evaluate the performance of the system.

3.2 Experimental Settings

In our CSECU-DSG system, we use the hashtag, URL, and username stripping techniques for tweets. We utilize the Huggingface (Wolf et al., 2019) transformer model RoBERTa (Liu et al., 2019) and fine-tune it with PyTorch (Paszke et al., 2019). Using the average of the top four layers, a 768-dimensional document embedding is generated from RoBERTa, which is forwarded to two stacked layers of the LSTM module and a fully connected linear architecture, yielding 1024- and 512-dimensional feature vectors, respectively. A 1536-dimensional fusion vector is generated by concatenating the output vector of the stacked LSTM and the linear architecture. Another feed-forward fully connected linear layer produces output from the 1536-dimensional fusion vector.

Parameter & Embeddings	Settings	
	Task 3a	Task 3b
learning_rate	1e-5	2e-5
max_epoch	15	4
batch_size	4	4
anneal_factor	0.5	0.5
patience	3	2

Transformer document embeddings	"roberta-base",
	layers = "-1,-2,-3,-4",
	layer_mean = True	

LSTM	input_size = 768,
	hidden_size = 1024,
	num_layers = 2,
	bidirectional = False	

Linear architecture	input_size = 768,
	output_size = 512	

Table 3: Model configuration and settings.

We use the FLAIR (Akbik et al., 2019) framework to implement our system. We train our system with the provided training data. We use Google Colab's GPU and set the set_seed = 42 to generate the reproducible results. We experiment with epochs in range [4,8,12,15] with different values of patience, batch size in range [4,8], learning rate in range [1e-5, 2e-5, 2e-4, 3e-4]. For the embedding settings, we consider the last layer, the average of the last four layers in RoBERTa, LSTM hidden_size 256,512,1024, LSTM num_layers 1,2,3, and linear layer feature value 256,512. Table 3 describes the set of optimal parameters used to design our proposed model. Default settings are used for the other parameters.

3.3 Results and Analysis

Category	F1-Score	Precision	Recall
Task 3a: tweet classification			
Dev set	0.6320	0.6489	0.6159
Test set	0.6082	0.6555	0.5673
Test (Median)	0.5859	0.6170	0.5577
Task 3b: WebMD review classification			
Dev set	0.9031	0.8767	0.9312
Test set	0.8608	0.8472	0.8748
Test (Median)	0.8432	0.8436	0.8646

Table 4: Results on the development and test sets.

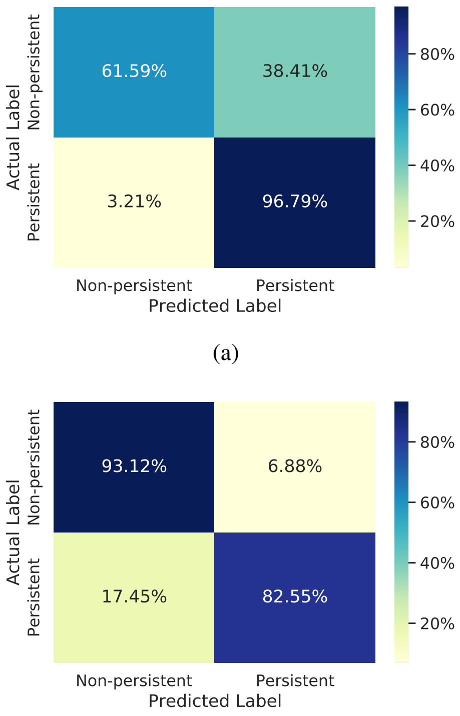
Table 4 shows the performance of our best-performing system based on the development set and the official results on test data in the individual sub-task. Compared to the official median scores on the test set computed using the participants' submissions, our system scored 61% and 86% F1-scores in subtasks 3a and 3b, respectively. With an 86% F1-score on WebMD review classification, we ranked first at task 3b.

3.4 Discussion

To qualitatively analyze the effectiveness of components utilized in our system, we perform an ablation study on the development set of task 3a and 3b. The experimental results are summarized in Table 5. The experimental result shows that the RoBERTa model performs pretty well, but combining RoBERTa with the LSTM layer increases the F1-score whereas incorporating stacked LSTM with RoBERTa brings a shrink in the F1-score. RoBERTa with LSTM and fully connected layer provides 3.37% and 0.22% growth in F1-score, in contrast, our proposed unified approach of RoBERTa, stacked LSTM, and fully connected linear architecture lead to a rise of 5.53% and 1.45% in F1-score on task 3a and 3b, respectively. Though RoBERTa with stacked LSTM lessens the F1-score, competitive performance appears for RoBERTa with stacked LSTM and fully connected layer inferring the efficacy of fusion of stacked LSTM and fully connected layer.

Method	Task 3a: tweet data			Task 3b: WebMD review		
	F1-score	Precision	Recall	F1-score	Precision	Recall
RoBERTa	0.5785	0.6731	0.5072	0.8886	0.8735	0.9042
RoBERTa + LSTM	0.5847	**0.7041**	0.5000	0.8903	0.8670	0.9150
RoBERTa + Stacked LSTM	0.5283	0.5512	0.5072	0.8830	0.8702	0.8961
RoBERTa + LSTM + Fully Connected Layer	0.6122	0.7009	0.5435	0.8908	**0.8896**	0.8920
Proposed Method: RoBERTa + Stacked LSTM + Fully Connected Layer	**0.6320**	0.6489	**0.6159**	**0.9031**	0.8767	**0.9312**

Table 5: Ablation study on the development set of task 3a and 3b. The best results are highlighted in boldface.



(a)

(b)

Figure 2: Confusion matrix of task 3a and 3b.

Tweet/Review	Actual	Predicted
Task 3a: tweet classification		
E#1: So for all your information, no I did not take zofran when I was pregnant.	0	1
E#2: I'll be back on Lipitor tomorrow, I'm sure.	1	0
Task 3b: WebMD review classification		
E#3: I had several debillatating side effects .	0	1
E#4: This med is useless for pain. The only pain meds that works is hydrocodone.	1	0

Table 6: Erroneous prediction of proposed method.

4 Error Analysis

Figure 2 shows the confusion matrix for subtasks 3a and 3b on the development set. We observe a relatively high proportion of misclassified non-persistent tweets, while the system performs fairly well in detecting non-persistent reviews.

Further, we articulate some erroneous predictions of our system in Table 6 to look into the reasons for misclassification. The first example (E#1) expresses a person's loftiness for not taking medicine, but our system misinterprets it as non-adherence to medication. The second example (E#2) implicitly conveys a patient's intention to skip medication dose today and stick to it from tomorrow. Due to the implicit nature of meaning, our system fails to detect the non-adherence characteristics of the tweet. On the contrary, for WebMD review classification, sample E#3 indicates aftereffects of consuming a drug that is misapprehended by our system. The first part of the sample E#4 shows the inefficacy of a drug whereas the second part reveals the effectiveness of another drug. Owing to the contradictory meaning, our system fails to classify the review correctly. Besides, the dataset is quite imbalanced as depicted in Table 2. Therefore, an appropriate strategy to handle implicitness, brevity, ambiguity, and unusual structure of text will improve the performance of our system.

5 Conclusion

In this paper, we introduce an approach to identify the self-declared drug-changing information by integrating RoBERTa, LSTM, and fully connected linear layers. In the future, we intend to explore biomedical-based transformers and model ensembling to distill better feature representation.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

Weissenbacher Davy, Z. Klein Ari, Gascó Luis, Estrada-Zavala Darryl, Krallinger Martin, Guo Yuting, Ge Yao, Sarker Abeed, Lucia Schmidt Ana, Rodriguez-Esteban Raul, Leddin Mathias, Magge Arjun, M. Banda Juan, Davydova Vera, Tutubalina Elena, and Gonzalez-Hernandez Graciela. 2022. Overview of the seventh social media mining for health applications #smm4h shared tasks at coling 2022. In *Proceedings of the Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Ralf C Staudemeyer and Eric Rothstein Morris. 2019. Understanding lstm–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.