NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001307

STIS at the NTCIR-17 MedNLP-SC Task: Incorporating Sentiment to Transformer Architecture for Adverse Drug Event Detection on Social Media

Lya Hulliyyatus Suadaa
Department of Statistical Computing
Politeknik Statistika STIS
lya@stis.ac.id

Eko Putra Wahyuddin
Department of Statistical Computing
Politeknik Statistika STIS
ekoputra@stis.ac.id

Farid Ridho
Department of Statistical Computing
Politeknik Statistika STIS
faridr@stis.ac.id

ABSTRACT

This paper presents the system and results of the STIS team for the Social Media (English) subtasks of the NTCIR-17 MedNLP-SC Task. We proposed incorporating the sentiment of social media texts into a pre-trained Transformer model in detecting adverse drug events on social media. A lexicon-based and rule-based sentiment analysis VADER model was used to predict each tweet sentiment. Based on the experimental results of the ADE vs. non-ADE binary classification task, our proposed fine-tuned model outperformed the baseline by a slight difference. Specifically, our model achieves a better F1 score for 9 of 22 symptoms in the symptom detection task.

TEAM NAME

STIS

SUBTASKS

SM-ADE-EN

KEYWORDS

sentiment, pre-trained Transformers, adverse drug

1 INTRODUCTION

Billions of people worldwide use social media to make networking and share information. Related to product reviews, consumers often use social media to share their product testimonials. Specifically in health fields, drug-related posts on social media could be a potential source for seeking information about drug safety.

The STIS team participated in the Social Media (English) subtasks of the NTCIR-17 MedNLP-SC Task [11]. In this subtask, we must detect adverse drug events on social media texts. Our proposed approach utilizes tweet sentiment while detecting symptoms in social media texts. We assumed that negative sentiment should indicate the symptom's availability in the texts. In this paper, we predicted the sentiment of the texts and incorporated the sentiments into our Transformer architecture in detecting adverse drug events.

2 RELATED WORK

Traditional statistical methods are not sufficient to predict future ADE patients [9]. Some methods are limited by the assumptions that must be fulfilled and the number of variables that can be used. Machine learning methods can overcome these limitations and produce more accurate predictions, as done by Yu et al [12]. They used various machine learning methods to predict the occurrence of ADEs in 1,746 pediatric inpatients in China. Data were obtained from electronic medical records. The best model they got had 44% precision, 25% recall, and 31.88% F1. Their model outperformed the results of Ji et al. [3], which only had a precision of 13.3% with almost the same study configuration.

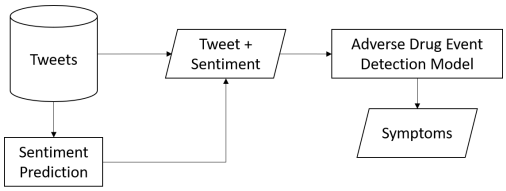

Figure 1: Proposed symptom detection architecture incorporating sentiments.

Still using electronic medical record data, Rebane et al [8] used a larger amount of data. They used 1,314,646 patient data obtained from the research infrastructure of the Swedish Health Record Research Bank, Health Bank at Stockholm University. The method used is also more advanced. They use deep learning with the RETAIN and RETAIN-ERF models. Both models use temporal attention mechanisms to include temporally highly relevant information about medical events. The best model is obtained by the RETAIN-ERF model with an F1 value of 83%.

The development of social media such as Twitter can provide additional information about ADE. Moh et al [5] extracted ADE events from 5839 tweets. The best model they obtained was the Support Vector Machine (SVM) model with an F1 score of 63%. Developments in natural language processing support the use of text data to predict ADE. Murphy et al [6] conducted a literature review of 29 articles that applied NLP to predict ADE. There are 17 articles that use named entity recognition and 15 articles that use classification models. In general, the performance of the model built is good, but it becomes worse if we only focus on the ADE case. This is because some non-ADE symptoms are inconsistent, for example "cough" could be ADE in the context of lisinopril, but an indicative in the context of codeine, or a symptom in the context of tuberculosis [6]. Therefore, we need a model that can understand the context of sentences better. One model that can memorize and understand the context better is Transformers.

Pre-trained Transformer models have been successfully applied in many downstream natural language processing tasks, including text classification. For adverse drug event detection tasks, Chaichulee et al. [1] tested different general-domain pre-trained BERT models,

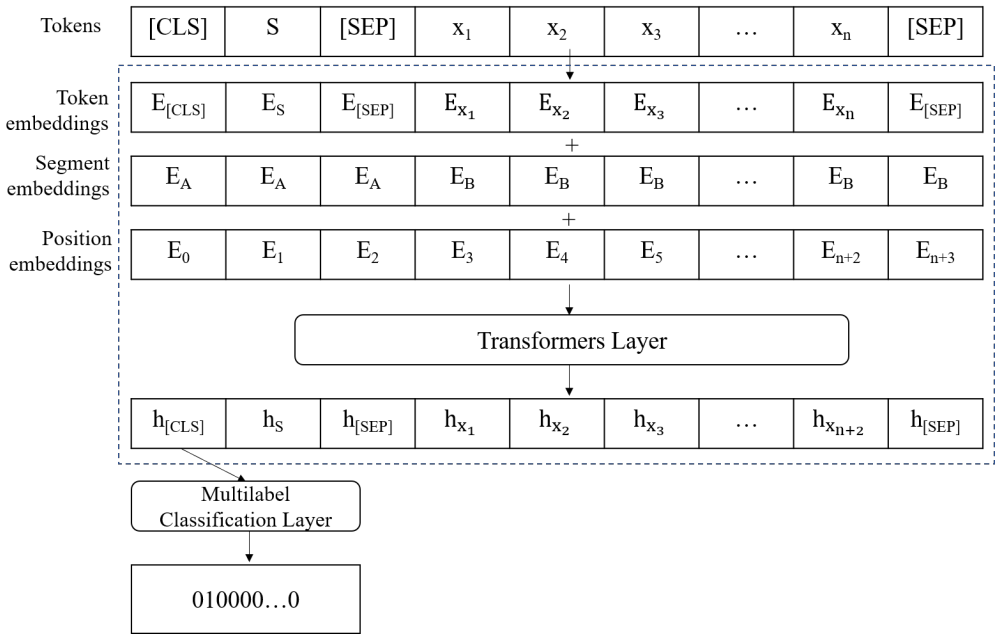

Figure 2: Proposed Transformers architecture incorporating sentiments for multilabel classification.

including BERT, mBERT, XLM-RoBERTa, and WanchanBERTa, and domain-specific AllergyRoBERTa in fine-tuning the models using multilabel classification approach. Overall, a plain BERT models had the highest performance.

3 METHODS

Our proposed idea is to incorporate the tweet's sentiment as input when predicting symptoms in the adverse drug detection task. We assumed that symptoms caused by adverse drugs should be extracted from tweets with negative sentiment. The architecture of the proposed approach can be seen in Figure 1.

3.1 Datasets

The datasets used in this paper consist of tweets containing Adverse Drug Events (ADEs) vs. those not containing ADEs which are manually labeled by annotators [11]. The 22 symptom labels in the tweets containing ADEs are "nausea," diarrhea," "fatigue," "vomiting," "loss of appetite," "headache," "fever," "interstitial lung disease," "liver damage," "dizziness," "pain," "alopecia," "analgesic asthma syndrome," "renal impairment," "hypersensitivity," "insomnia," "constipation," "bone marrow dysfunction," "abdominal pain," "hemorrhagic cystitis," "rash," "stomatitis."

Table 1: Number of tweets for each category

Category	Train	Validation	Test	Total
Positive	2462	618	817	3897
Neutral	930	252	278	1460
Negative	2979	723	898	4600
Total	6371	1593	1993	9957

Table 2: Results of ADE vs non-ADE Binary Classification from NTCIR-17 MedNLP-SC. The best score is in bold.

Metrics		BERT Baseline	BERT Sentiment
Precision	ADE	0.71	**0.75**
	Macro Average	0.82	**0.83**
Recall	ADE	**0.80**	0.78
	Macro Average	0.84	0.84
F1	ADE	0.76	**0.77**
	Macro Average	0.82	**0.84**

3.2 Sentiment Prediction

In this study, the Valence Aware Dictionary for Sentiment Reasoning (VADER) model was used to perform sentiment analysis. VADER is a lexicon-based and rule-based sentiment analysis model that is specifically adapted to the sentiments expressed on social media [4]. VADER will produce a sentiment score in the range of -1 to 1. The closer to 1, a text is considered to have a more positive sentiment and vice versa.

VADER can handle words, abbreviations, slang, emoticons, and emojis commonly found on social media. The performance of VADER is demonstrated directly in the original scientific article [4] by comparing it with other models. The results shown are that VADER is better than other models, especially when applied to comment data on social media. Therefore, VADER is often used for sentiment analysis on social media data, especially Twitter data. For example, Pano and Kashef [7] used VADER to analyze Tweets to predict bitcoin prices. Likewise, Singh et al [10] conducted sentiment analysis using VADER to predict stock prices.

Table 3: Results for each symptom from NTCIR-17 MedNLP-SC. The best score is in bold.

Symptom	Precision		Recall		F1		Number of Samples
	BERT Baseline	BERT Sentiment	BERT Baseline	BERT Sentiment	BERT Baseline	BERT Sentiment	
Nausea	0.74	**0.81**	0.89	**0.92**	0.81	**0.86**	120
Diarrhea	0.72	**0.78**	**0.79**	0.76	0.76	**0.77**	136
Fatigue	0.73	0.73	0.79	**0.84**	0.76	**0.78**	56
Vomiting	0.77	0.77	0.91	0.91	0.83	0.83	22
Loss of appetite	0.66	**0.81**	0.81	**0.85**	0.72	**0.83**	52
Headache	0.76	**0.87**	0.89	**0.91**	0.82	**0.89**	57
Fever	**0.57**	0.52	**0.72**	0.60	**0.63**	0.56	53
Interstitial lung disease	**0.67**	0.00	**1.00**	0.00	**0.80**	0.00	2
Liver damage	0.00	0.00	0.00	0.00	0.00	0.00	2
Dizziness	0.56	**0.65**	0.77	**0.85**	0.65	**0.73**	13
Pain	0.56	**0.65**	0.44	**0.46**	0.50	**0.54**	72
Alopecia	0.88	**1.00**	**0.88**	0.50	**0.88**	0.67	8
Analgesic asthma syndrome	**1.00**	0.89	0.83	**0.89**	**0.91**	0.89	18
Renal impairment	**0.60**	0.00	**0.60**	0.00	**0.60**	0.00	5
Hypersensitivity	0.71	**0.74**	**0.79**	0.61	**0.75**	0.67	28
Insomnia	0.65	**0.80**	**0.32**	0.24	**0.43**	0.36	34
Constipation	0.74	**0.81**	**0.90**	0.68	**0.81**	0.74	31
Bone marrow dysfunction	0.00	0.00	0.00	0.00	0.00	0.00	2
Abdominal pain	0.61	**0.73**	**0.83**	0.82	0.71	**0.77**	88
Hemorrhagic cystitis	**0.75**	0.00	**0.75**	0.00	**0.75**	0.00	4
Rash	0.68	**0.74**	**0.79**	0.61	**0.73**	0.67	33
Stomatitis	0.61	**0.90**	**0.50**	0.41	0.55	**0.56**	22

In this study, we preprocess each tweet by changing the characters to lowercase. After that, each tweet is processed by the VADER model and produces a sentiment score. Any tweet that has a score less than -0.05 will be categorized as negative, a score greater than 0.05 will be categorized as positive and otherwise will be categorized as neutral. Table 1 shows the number of tweets for each category. Overall there were 4600 negative tweets, 1460 neutral tweets, and 3897 positive tweets.

3.3 Transformer-based Adverse Drug Event Detection

Incorporating sentiment predictions from previous step, we finetune pre-trained Transformer BERT for adverse drug event detection task. Adopting the finetuned BERT approach in question answering task [2], we preprocess sentiment and tweet tokens as inputs by inserting two special tokens, [CLS] and [SEP]. The [CLS] token is added to the beginning of input, and the [SEP] token is inserted after the sentiment token to separate the sentiment and tweet segments. The [SEP] token is also added to the end of input.

Three types of embedding are assigned to each input: token embeddings representing the semantic of tokens, segment embeddings marking different types of a sequence of tokens, and positional embeddings indicating the token position within the sequences. We used the token representations from the top hidden layer ($h_{[CLS]}$) as context embeddings. The hidden state was obtained as follow,

$$h_{[CLS]} = \text{Transformer}([CLS] + S + [SEP] + X + [SEP]), \quad (1)$$

where $S \in$ {positive, neutral, negative} and $X = (x_1, x_2, ..., x_n)$ is a sequence of tweet tokens.

Then $h_{[CLS]}$ were fed to the multilabel classification layer to obtain the availability of symptoms in each tweet. Our proposed Transformers architecture is shown in Figure 2.

4 EXPERIMENTS

The overall performances of our proposed approach using pretrained Transformers are shown in Table 2. We can see that our approach by incorporating sentiment in the Finetuned BERT outperformed the baseline by a slight difference. To predict ADE, our proposed model has an F1 score of 0.01 higher than the baseline model. The macro F1 value is also 0.02 higher than the baseline. It shows that adding sentiment as additional information to BERT can improve model performance when predicting ADE cases and non-ADE cases. Based on the recall values obtained, the baseline model can correctly predict 80% of ADE cases. Meanwhile, our proposed model was only able to correctly predict 78% of ADE cases. However, the 2% decrease in recall was offset by a 4% increase in precision value from 71% to 75%.

Table 3 displays the precision, recall, and F1 values for each type of symptom. Based on the F1 score, out of 22 symptoms, our model is better than the baseline model in predicting nine symptoms, namely nausea, diarrhea, fatigue, loss of appetite, headache, dizziness, pain, abdominal pain, and stomatitis. There are three symptoms that our model performs the same as baseline, namely vomiting, bone marrow dysfunction, and liver damage. For the other ten symptoms, the baseline model is better than our proposed model. It can be

Table 4: Samples of symptom.

Tweets	Sentiment	Predicted Symtomp without Sentiment	Predicted Symtomp with Sentiment
I thought I was okay because I took all the antibiotics and my fever is down and I'm eating, but I had a fever during the night. I wonder if I can sleep.	negative	none	fever
Mesalazine Enteric Tablets 250mg Day 324The hardness of the stool is a little better than yesterday, but the abdominal pain is still there...I feel hungry and not down, so we'll see.	negative	none	abdominal

concluded that when the number of training samples is not that small, overall, the proposed system obtains higher precision and lower recall. A possible hypothesis is that the system might tend to require a negative sentiment to predict an ADE, therefore refraining from predicting ADEs for non-negative tweets.

The samples of symptom prediction for several tweets using the finetuned model in the NTCIR-17 testing dataset are shown in Table 4. In some cases, incorporating sentiment as additional tweet information could help the model detect the correct symptoms.

5 CONCLUSIONS

In this paper, we proposed incorporating the sentiment of social media texts into a pre-trained Transformer model in detecting adverse drug events on social media. We used VADER, a lexicon-based and rule-based sentiment analysis model, to perform sentiment analysis. Then, we concatenated a pair of sentiments and tweets as input and fine-tuned the pre-trained Transformer BERT model for the adverse drug event detection task. Based on the experimental results of the ADE vs. non-ADE binary classification task, our proposed fine-tuned model outperformed the baseline by a slight difference. Specifically, our model achieves a better F1 score for 9 of 22 symptoms in the symptom detection task. These results could be a good sign of the sentiment effect in the improved model.

REFERENCES

[1] Sitthichok Chaichulee, Chissanupong Promchai, Tanyamai Kaewkomon, Chanon Kongkamol, Thammasin Ingviya, and Pasuree Sangsupawanich. 2022. Multi-label classification of symptom terms from free-text bilingual adverse drug reaction reports using natural language processing. *PLOS ONE* 17, 8 (08 2022), 1–22. https://doi.org/10.1371/journal.pone.0270595

[2] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[3] Huan huan Ji, Lin Song, Jian wen Xiao, Yu xia Guo, Ping Wei, Ting ting Tang, Xiao jiang Tian, Xue wen Tang, and Yun tao Jia. 2018. Adverse drug events in Chinese pediatric inpatients and associated risk factors: a retrospective review using the Global Trigger Tool. *Scientific Reports* 8, 1 (Feb. 2018). https://doi.org/10.1038/s41598-018-20868-2

[4] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

[5] Melody Moh, Teng-Sheng Moh, Yang Peng, and Liang Wu. 2017. On adverse drug event extractions using twitter sentiment analysis. *Network Modeling Analysis in Health Informatics and Bioinformatics* 6, 1 (Sept. 2017). https://doi.org/10.1007/s13721-017-0159-4

[6] Rachel M. Murphy, Joanna E. Klopotowska, Nicolette F. de Keizer, Kitty J. Jager, Jan Hendrik Leopold, Dave A. Dongelmans, Ameen Abu-Hanna, and Martijn C. Schut. 2023. Adverse drug event detection using natural language processing: A scoping review of supervised learning methods. *PLOS ONE* 18, 1 (Jan. 2023), e0279842. https://doi.org/10.1371/journal.pone.0279842

[7] Toni Pano and Rasha Kashef. 2020. A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19. *Big Data and Cognitive Computing* 4, 4 (Nov. 2020), 33. https://doi.org/10.3390/bdcc4040033

[8] Jonathan Rebane, Isak Samsten, and Panagiotis Papapetrou. 2020. Exploiting complex medical data with interpretable deep learning for adverse drug event prediction. *Artificial Intelligence in Medicine* 109 (Sept. 2020), 101942. https://doi.org/10.1016/j.artmed.2020.101942

[9] Mio Sakuma, David W. Bates, and Takeshi Morimoto. 2012. Clinical prediction rule to identify high-risk inpatients for adverse drug events: the JADE Study. *Pharmacoepidemiology and Drug Safety* 21, 11 (Aug. 2012), 1221–1226. https://doi.org/10.1002/pds.3331

[10] Ashutosh Singh, Jyotsna Singh, and Anindya Ghosal. 2022. Impact Of Social Media On Stock Market- A Case Of Sentiment Analysis Using Vader. In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*. IEEE. https://doi.org/10.1109/com-it-con54601.2022.9850668

[11] Shoko Wakamiya, Lis Kanashiro Pereira, Lisa Raithel, Hui-Syuan Yeh, Peitao Han, Seiji Shimizu, Tomohiro Nishiyama, Gabriel Herman Bernardim Andrade, Noriki Nishida, Hiroki Teranishi, Narumi Tokunaga, Philippe Thomas, Roland Roller, Pierre Zweigenbaum, Yuji Matsumoto, Akiko Aizawa, Sebastian Möller, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Shuntaro Yada, and Eiji Aramaki. 2023. NTCIR-17 MedNLP-SC Social Media Adverse Drug Event Detection: Subtask Overview. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17*. National Institute of Informatics (NII). https://doi.org/10.20736/0002001327

[12] Ze Yu, Huanhuan Ji, Jianwen Xiao, Ping Wei, Lin Song, Tingting Tang, Xin Hao, Jinyuan Zhang, Qiaona Qi, Yuchen Zhou, Fei Gao, and Yuntao Jia. 2021. Predicting Adverse Drug Events in Chinese Pediatric Inpatients With the Associated Risk Factors: A Machine Learning Study. *Frontiers in Pharmacology* 12 (April 2021). https://doi.org/10.3389/fphar.2021.659099