JAMIA Open, 4(3), 2021, 1–9
doi: 10.1093/jamiaopen/ooab081
Research and Applications

Research and Applications

Deep learning models in detection of dietary supplement adverse event signals from Twitter

Yefeng Wang [1], Yunpeng Zhao [2], Dalton Schutte [1,3], Jiang Bian [2], and Rui Zhang [1,3]

[1]Institute for Health Informatics, University of Minnesota, Minneapolis, Minnesota, USA, [2]Department of Health Outcomes & Biomedical Informatics, University of Florida, Gainesville, Florida, USA, and [3]Department of Pharmaceutical Care & Health Systems, University of Minnesota, Minneapolis, Minnesota, USA

Corresponding Author: Rui Zhang, PhD, Institute for Health Informatics, Department of Pharmaceutical Care & Health Systems, University of Minnesota, 516 Delaware St SE, Minneapolis, MN 5545, USA; zhan1386@umn.edu

Received 11 May 2021; Revised 30 August 2021; Editorial Decision 2 September 2021; Accepted 7 September 2021

ABSTRACT

Objective: The objective of this study is to develop a deep learning pipeline to detect signals on dietary supplement-related adverse events (DS AEs) from Twitter.
Materials and Methods: We obtained 247 807 tweets ranging from 2012 to 2018 that mentioned both DS and AE. We designed a tailor-made annotation guideline for DS AEs and annotated biomedical entities and relations on 2000 tweets. For the concept extraction task, we fine-tuned and compared the performance of BioClinical-BERT, PubMedBERT, ELECTRA, RoBERTa, and DeBERTa models with a CRF classifier. For the relation extraction task, we fine-tuned and compared BERT models to BioClinical-BERT, PubMedBERT, RoBERTa, and DeBERTa models. We chose the best-performing models in each task to assemble an end-to-end deep learning pipeline to detect DS AE signals and compared the results to the known DS AEs from a DS knowledge base (ie, iDISK).
Results: DeBERTa-CRF model outperformed other models in the concept extraction task, scoring a lenient microaveraged F1 score of 0.866. RoBERTa model outperformed other models in the relation extraction task, scoring a lenient microaveraged F1 score of 0.788. The end-to-end pipeline built on these 2 models was able to extract DS indication and DS AEs with a lenient microaveraged F1 score of 0.666.
Conclusion: We have developed a deep learning pipeline that can detect DS AE signals from Twitter. We have found DS AEs that were not recorded in an existing knowledge base (iDISK) and our proposed pipeline can as sist DS AE pharmacovigilance.

Key words: dietary supplements, social media, adverse events, deep learning, natural language processing

LAY SUMMARY

This study has developed a deep learning-based natural language processing pipeline to identify dietary supplements (DS) adverse events (AE) from Twitter. The pipeline is consisted of 2 modules, one for identifying the word or phrases that corresponds to DS and AE symptoms, another for extracting the relation between DS and AE. The pipeline was able to find 3791 DS AE pairs, 1563 DS deficiency AE pairs (where the AE is caused by lacking intake of a certain DS), and 16 222 DS indication pairs. The DS AE signals detected from Twitter have a small overlap with the existing DS knowledge base.

INTRODUCTION

Dietary supplements (DSs) are gaining popularity depicted by their steady escalating usage which reaches all-time high in 2019 according to an annual survey on consumers' DS usage conducted by the Council for Responsible Nutrition (CRN). Seventy-seven percent of Americans have used at least one DS, and adults between the ages 35 and 54 have the highest usage of DS.[1] However, DS regulatory policies are different and less rigorous than those covering their drug counterpart. As per the Dietary Supplement Health and Education Act of 1994 (DSHEA),[2] both DS products and DS ingredients are regulated by the Food and Drug Administration (FDA), but clinical approval trials for DS safety and efficacy are not mandatory.[3] As a result, there is an estimated over 23 000 emergency department visits per year were attributed to DS use.[4]

The existing pharmacovigilance infrastructure around DS primarily relies on postmarketing spontaneous reporting system (SRS) where DS manufacturers, researchers, clinicians, and consumers voluntarily report adverse events (AEs) online. In the United States, the Center for Food Safety and Applied Nutrition (CFSAN) under the FDA launched the CFSAN Adverse Event Reporting System (CAERS) in 2003 to facilitate postmarket monitoring and surveillance of adverse event reports (AERs) associated with food, cosmetic, and DS.[5] The primary purpose of CAERS is to enhance consumer safety through the real-time assessment of AERs.[6] However, the distribution of the AER reporting sources is heavily skewed.[7] By Q2 of 2016, healthcare professionals had contributed 237 996 AERs, while DS consumers only contributed only 69 267 AERs to the SRS.[8] It is possible that the consumers may not know how to use the SRS to report the AEs they experienced, or they might not be aware that such systems exist. Moreover, when health professionals report adverse events (AEs), they focus more on serious AEs that may present grave danger to patient health but overlook more common AEs that are not life-threatening.[9] Therefore, additional data sources that put more emphasis on DS consumers are necessary for effective DS surveillance and monitoring.

Social media (SM) has emerged as a valuable resource for pharmacovigilance due to the accessibility and the timeliness of its data.[10,11] The sheer quantity of self-reports of drug AEs on SM implies that it is an indispensable supplementing resource for SRS.[12] This is especially true after the COVID-19 lockdown as people are getting more used to obtain and exchange health-related information over SM platforms. Golder et al[13] have found that SM data adequately represent DS AEs of mild symptoms, which complements the reporting bias of SRS toward more serious conditions. Duh et al[14] have found that drug AEs were reported on SM 11 months earlier on average than other platforms such as an SRS.

The identification of DS AEs from SM data can be summarized into 2 tasks: (1) concept extraction, where the terms that correspond to DSs and AEs are identified; (2) relation extraction, where the relations between these terms are identified. These 2 tasks can be done either stepwise or jointly.[15] Various corpora have been developed to train machine learning models and benchmark their performances on both tasks, such as the ones from Integrating Biology and the Bedside (i2b2),[16] ShARe/CLEF,[17] SemEval challenges,[18] and Social Media Mining for Health (SMM4H) shared task.[19] However, the corpora and the models trained on them put more emphasis on prescription drugs rather than DS. Compared to drug AE identification, DS AE signal detection has its own unique challenges. Unlike prescription drugs, DS concepts have larger variations.[20] Some DS products are of a single ingredient, for example, a Vitamin C tablet. But some DS products can be a food where the DS exists as an active ingredient. Therefore, our proposed model should be able to identify DS concepts from a mixture of DS ingredient and food concepts. Twitter users are more likely to post about how DS improves an existing condition than causing AEs. The low signal-to-noise ratio of DS AE signals from Twitter data was another challenge to our proposed model. Currently, to the best of our knowledge, while there are many pipelines that extract and identify drug AEs, there is no existing pipeline to extract and identify DS AEs from Tweets. Therefore, our study would like to bridge the gap in DS AE signal detection by (1) constructing a manually annotated tweet dataset regarding the DS AE and (2) develop a transformer-based natural language processing (NLP) pipeline to automatically extract DS AEs from tweets.

MATERIALS AND METHODS

The overview of the methods was shown in Figure 1. We collected and annotated a set of DS-related tweets and compared the performance of traditional embeddings with contextual embeddings in the concept extraction and relation extraction tasks, respectively. The best-performing models were used to assemble an end-to-end pipeline for the identification of DS AEs from tweets. We compared the signals generated by the pipeline on a larger corpus with an existing DS knowledge base (ie, integrated dietary supplement knowledge base [iDISK]).[21]

Data collection

To retrieve the tweets with co-occurrence of DS and symptom/body organ terms, we compiled 2 lists of terms. The DS term list was obtained from our previous study, which contains 332 DS terms including 31 commonly used DS names and their name variants.[23] The symptom/body organ term list contains 14 143 terms by integrating the ADR lexicon[24] and the iDISK knowledge base. We selected only English tweets to develop the concept extraction and relation extraction models. A total of 247 807 tweets that satisfy the criteria were found from a Twitter database we constructed in prior work[25] using the Twitter streaming application programming interface, covering daily public tweets from 2012 to 2018.

Data preprocessing

We employed the *ekphrasis* package[26] to remove uniform resource locators, user handle (eg, @username), hashtag symbol ("#"), and emoji characters. Contractions such as "*doesn't*," "*won't*" were expanded into "*does not*" and "*will not*," respectively. Hashtags were segmented into their constituent words (eg, "*ILoveVitaminC*" would be segmented into "*I Love Vitamin C*"). Stop words were not removed, because some of the stop words are meaningful (eg, "*throw up*" is a common phrase to describe vomiting, but after removing the stop words it would become "*throw*," losing the original meaning).

Annotation

We randomly selected 2000 tweets from the dataset for annotation. Two annotators manually reviewed the tweets, highlighting all biomedical entities and relations between them with brat annotation tool. Initially, a random sample of 100 tweets was selected for the creation of annotation guideline and calculation of inter-rater agree-

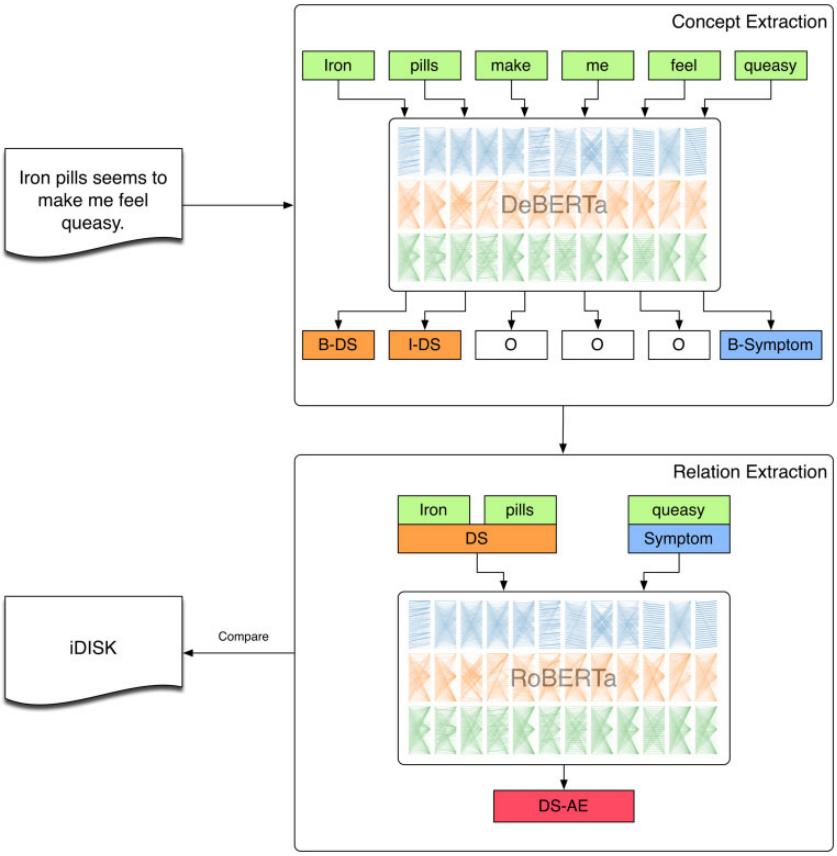

Figure 1. The overview of our study workflow.[22] Here, we have shown the process of a tweet containing a DS AE going through the pipeline with the best-performing concept extraction and relation extraction model. The figure within the RoBERTa model box has demonstrated the self-attention within the transformer layers with respect to the input tweet. To compare the performance of different models, we will switch the models for concept extraction and relation extraction module accordingly. DS AE: dietary supplement-related adverse events.

ment. We describe the detailed annotation guideline in the following subsection.

Concept extraction

The Beginning-Inside-Outside representation was selected to label the entities. A concrete annotation example is given in Figure 2. Note that O tags were not highlighted in this example.

We defined 3 entity types: DS, symptoms, and body organs. Supplements include both oral DS and supplements that are topically applied. For example, Vitamin E taken orally, and Vitamin E oil used on skin should all be included. The form of supplement is kept as a part of the named entity. For example, the word "*oil*" in "*oregano oil*" should not be omitted. For herbal supplement, it is common that the supplement is a part of the plant. The word that describes the part is also kept within the named entity. For example, the word "*seed*" needs to be annotated in the noun phrase "*grape seed*." We annotated the deficiency of the supplement, for example, "*vitamin b12 deficiency increases risk of cancer*," as "*deficiency*." With the deficiency information available, the machine learning model can avoid making the mistake of identifying cancer as an AE of vitamin B12 in this example.

Symptoms are entities that describe the specifics of a DS AE, such as "*cold*," "*cough*," "*diarrhea*," "*cancer*," "*throw up*," "*feel sick*," etc. However, some tweets did not specify the symptoms of a DS AE. For example, in the tweet "*too much vitamin A is bad for your liver*," although no symptoms were mentioned, the body organs where the supplement might take effect could indicate a DS AE. Therefore, we would annotate body organ entities as a DS AE signal as well.

Relation extraction

We defined 2 binary relations based on the above definition of entities: indication and adverse events. Purposes are positive effects due to the use of the supplements. For example, "*Vitamin D reduces fatigue*" implies that the purpose of using Vitamin D is to deal with fatigue. "*Kava kava balances mood*" indicates that the purpose of using "*kava kava*" is to balance mood. AEs are negative effects due to the use of the supplements. In the example of "*Excess vitamin D weakens bone*," Vitamin D has an undesired effect on bone. Iron has resulted an AE—queasy in the "*Iron pills seems to make me feel queasy*."

The 2000 fully annotated tweets contained 2244 DS entities, 2003 symptom entities, and 287 body organ entities. There are 1471 indication mentions and 442 AE mentions. The inter-rater agreement (kappa score) for the concept extraction task is 0.9416 and 0.8299 for the relation extraction task.

Concept extraction models

The performance of concept extraction is dependent on the word representation.[27] Traditional word embeddings such as Word2-Vec,[28] GloVe,[29] and fastText[30] managed to integrate word context information into a single vector. This presents an obstacle to

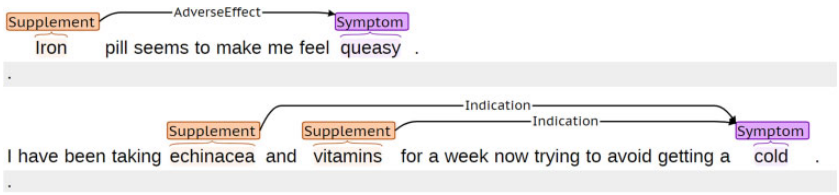

Figure 2. Two annotation examples. One corresponds to a DS AE annotation, another corresponds to a DS indication annotation. DS AE: dietary supplement-related adverse events.

the concept extraction task and the relation extraction task as the single vector cannot cope with polysemy. For example, the word "*cut*" can either be a noun meaning "*a flesh wound caused by a sharp object*," or be a verb referring to the action of "*penetrate with an edged instrument*," or be a part of a verbal phrase such as "*cut back*" which means "*to reduce*." However, Word2Vec, GloVe, and fastText could only learn a final word vector for the word "cut" regardless of its meaning in the text. It would be hard to isolate the "*flesh wound*" meaning of the word from the other ones if the same word vector were used for all future predictions. Contextual embeddings such as ELMo (Embeddings from Language Models)[31] and BERT (Bidirectional Encoder Representations from Transformers)[32] that can alter dynamically based on its meaning in context were designed to overcome such drawback. Instead of a final real-valued vector, contextual embeddings use all layers of a deep neural network trained on a large corpus. BERT employs a fine-tuning approach, where the pretrained deep neural network is further optimized with respect to the downstream tasks. BERT models have shown superior performance than traditional word embeddings in NLP tasks such as concept extraction, relation extraction, and question answering tasks.[33] BERT models have now become the major workforce in biomedical information extraction, and they have been extensively applied to drug AE classification, extraction, and normalization in SM data.[34–37] However, these state-of-the-art models have not been applied to DS AE signals detection to the best of our knowledge. In this study, we used the base BERT model as our baseline. We employed several approaches to improve the performance: (1) we substituted the softmax classifier with a CRF classifier downstream; (2) we used BERT models that were pretrained on biomedical corpora such as PubMed and PMC articles instead of general domain texts (Bioclinical-BERT[38] and PubMedBERT[39]); and (3) we tried more sophisticated architecture such as RoBERTa[40] (which uses GPT-2 byte-pair encoding), BioELECTRA[41] (which includes a generative adversarial network [GAN]-like component that generates false token), and DeBERTa[42] (which uses a disentangled attention mechanism) .

In this task, we compared machine learning and deep learning models to recognize biomedical entities including DS names, AE terms, and body organs. Among 2000 tweets, we held out 20% as a test set for model-to-model evaluation comparison. The remaining 80% of the tweets was used in 5-fold cross-validation to determine the best learning rate of a concept extraction model, for it is the most crucial hyperparameter to the training process. Default values are used for other hyperparameters as suggested by the *transformers* package. All models were trained for 20 epochs. We employed the evaluation metric of n2c2 2018 Shared Task 2 to compare the performance of the models. The lenient matching microaveraged F1 score was used as the primary criteria to compare model performance.[43]

Relation extraction models

The concept extraction model helped find the position of biomedical entities within a tweet. The next step is to identify the relations between the biomedical entities, especially between the DS and AE entities. Multiple relations exist between a DS and an AE term. In the tweet "*sounds weird but it works because vitamin C helps with sore throat*," the relation between the DS "*vitamin C*" term and the AE term "*sore throat*" is an indication, that is, the supplement was used with the intention to treat a symptom. But in the tweet "*note to self if you are used to 250 mg of niacin jump up to 500 mg the niacin flush is so intense*," the relation between "*niacin*" and "*flush*" is an adverse event, that is, the supplement caused the symptom. The relation extraction model should be able to differentiate DS AEs from DS indications. To measure the performance of the relation extraction model alone, we developed and compared models which predict relations (i.e., "no relation," "indication," or "AE") between all possible pairs between DS and AE entities.

We still focus on BERT models due to its versatility and stable performance in both concept extraction and relation extraction tasks. We compared the performance of BERT, Bioclinical-BERT, PubMedBERT, RoBERTa, and DeBERTa models on relation extraction tasks based on the evaluation metric of n2c2 2018 Shared Task 2.

We have referenced the implementation of Refs.[44,45] to evaluate the models and calculated the mean and the standard deviation of the lenient matching microaveraged F1 score based on the 5-fold cross-validation results. The best-performing model in each task is integrated into an end-to-end pipeline.

Evaluation of extracted DS AEs against the iDISK

We applied the end-to-end model trained on the annotated tweets to the full dataset and compared the signal detected from the machine learning pipeline to DS AE recorded in iDISK.

We tallied the occurrences of DS AE pairs and DS indication pairs extracted by the end-to-end pipeline. Especially, since the concept extraction model extracted DS deficiency information, the DS AEs extracted by our proposed pipeline can be divided into 2 categories: (1) normal DS AEs, where an AE is caused by taking a supplement and (2) DS deficiency AEs, where an AE is caused due to lacking intake of a supplement. We selected 50 most frequently mentioned DS AEs, DS deficiency AEs, and DS indications, respectively. For every pair, we manually reviewed the underlying tweet and check if the pair was correctly referring to a DS AE, a DS deficiency AE, or a DS indication. Only DS AEs and DS indications were compared to iDISK as it did not record DS deficiency AEs.

RESULTS

Concept extraction task

Table 1 shows the performance comparison among all the methods and word embeddings used in this task.

The baseline BERT model surprisingly scored a lenient matching microaveraged F1 score of 0.842, which outperformed BERT-CRF, Bioclinical-BERT-CRF, PubMedBERT-CRF, and ELECTRA-CRF models. Although DS AE terminologies are highly correlated to the biomedical corpora, the BERT models pretrained on PubMed and PMC articles did not improve the baseline model but worsened the performance instead, which only scored 0.769 and 0.760, respectively. This could be attributed to the noisy and colloquial nature of the language used on Twitter that is rarely seen in academic literature. BERT models that employed a more granular representation of tokens, for example, RoBERTa with byte-pair encoding, and DeBERTa, which employed a more sophisticated attention mechanism, has improved the performance of the concept extraction model. DeBERTa model in conjunction with a CRF classifier has achieved the best performance of 0.866, as shown in Table 1.

The entity-level model performance was shown in Figure 3. The performance of extracting symptom and DS concepts was consistently the highest among all concept types, scoring a microaveraged F1 score above 0.8. However, there are 2 outliers, PubMedBERT-CRF and ELECTRA-CRF, in DS concept extraction. These 2 models only scored a microaveraged F1 score around 0.6. The performance of extracting organ and food concepts for all models was around 0.7. Again, PubMedBERT-CRF and ELECTRA-CRF models had lower performance than other BERT-based models. The performance of extracting DS deficient concepts was the lowest among all models, as shown in Figure 2 by their lower average and longer error bars.

Relation extraction task

Table 2 shows the performance of the relation extraction task.

The baseline BERT model scored a microaveraged F1 score of 0.73. Two BERT models that were pretrained on biomedical corpora, BioClinical-BERT and PubMedBERT, were only able to improve the baseline performance slightly. The performance did not improve at all in the case of BioClinical-BERT, whose microaveraged F1 score stayed at 0.73. Using more sophisticated architecture proved to be more practical in improving the relation extraction performance. RoBERTa with byte-pair encoding was the best-performing model, which scored a microaveraged F1 score of 0.79, followed by DeBERTa that uses a disentangled attention mechanism with a microaveraged F1 score of 0.78.

End-to-end pipeline

We chose DeBERTa model for concept extraction and RoBERTa model with entity headings features for relation extraction due to their outperformance over all other models in each task. We assembled them into an end-to-end pipeline, that is, the output of the concept extraction model will be directly used as inputs into the relation extraction model. It is expected that the error would propagate along the pipeline and thus lead to an overall lower F1 score, and the evaluation results shown in Table 3 confirmed our expectations.

Comparison to iDISK

We applied the DS AE identification pipeline to the tweets in the full dataset excluding the annotated dataset. The pipeline was able to find 442 227 possible relations from extracted concepts in 247 807 tweets. Among these possible relations, 3791 were DS AE pairs and 16 222 were DS indication pairs. Table 4 presented the most frequently mentioned DS AE pairs in our dataset, while Table 5 presents examples of frequently mentioned DS indication pairs. We compared these DS AE and DS indication pairs to the records in the iDISK and found both existing and novel DS AE and DS indication extracted by our proposed pipeline.

The pipeline was also able to find DS AEs that were caused by lacking a certain type of DS, which we will refer to as "DS deficiency AEs." The pipeline was able to find 1563 DS deficiency AEs, and the most frequently mentioned ones were reported in Table 6. These deficiency AEs are not included in iDISK and thus could be supplementary to the DS-related AE monitoring.

DISCUSSION

While many annotated corpora are readily available for drug AE extraction, they were not tailored for DS AE extraction. In this study, we demonstrated the feasibility of using Twitter as a complementary resource for DS AE surveillance. We thus created our own annotated tweet dataset to evaluate machine learning and deep learning models and develop an end-to-end DS AE signal detection pipeline. The evaluation results have shown that BERT-based models with more granular encodings and sophisticated architecture, that is, RoBERTa and DeBERTa, outperformed the baseline model in both concept extraction and relation extraction tasks. Surprisingly, BERT models that were trained on biomedical literature were not able to outperform the baseline model. This suggests that SM text uses different vernaculars from academic literature, as shown by the underperformance of PubMedBERT in the concept extraction task. However, PubMedBERT performed well in the relation extraction tasks. This suggests that the ways to express DS AE relations on SM should be quite similar to the ways it was expressed in biomedical literature. But RoBERTa and DeBERTa still outperformed, implying that the transformer architecture could be a more important factor in performance than the corpora that the BERT model was pretrained on. We also noticed that the performance of DS deficient concept extraction was lower than other concept types, which can be attributed

Table 1. Performance of concept extraction models on the held-out test set. The performance of the best model is highlighted in bold.

	Lenient matching			Strict matching		
Model	Precision	Recall	F1	Precision	Recall	F1
BERT	0.831 ± 0.006	0.853 ± 0.011	0.842 ± 0.008	0.779 ± 0.007	0.800 ± 0.010	0.790 ± 0.008
BERT + CRF	0.819 ± 0.006	0.842 ± 0.009	0.830 ± 0.003	0.767 ± 0.009	0.789 ± 0.007	0.778 ± 0.004
Bioclinical-Bert + CRF	0.820 ± 0.006	0.844 ± 0.005	0.832 ± 0.005	0.767 ± 0.006	0.789 ± 0.005	0.778 ± 0.005
PubMedBERT + CRF	0.842 ± 0.011	0.707 ± 0.009	0.769 ± 0.005	0.763 ± 0.011	0.641 ± 0.10	0.697 ± 0.007
ELECTRA + CRF	0.842 ± 0.008	0.693 ± 0.015	0.760 ± 0.009	0.746 ± 0.008	0.614 ± 0.015	0.673 ± 0.010
RoBERTa + CRF	0.846 ± 0.006	0.876 ± 0.009	0.860 ± 0.004	0.797 ± 0.006	0.825 ± 0.010	0.811 ± 0.007
DeBERTa + CRF	**0.856 ± 0.009**	**0.876 ± 0.010**	**0.866 ± 0.003**	**0.812 ± 0.012**	**0.832 ± 0.005**	**0.822 ± 0.005**

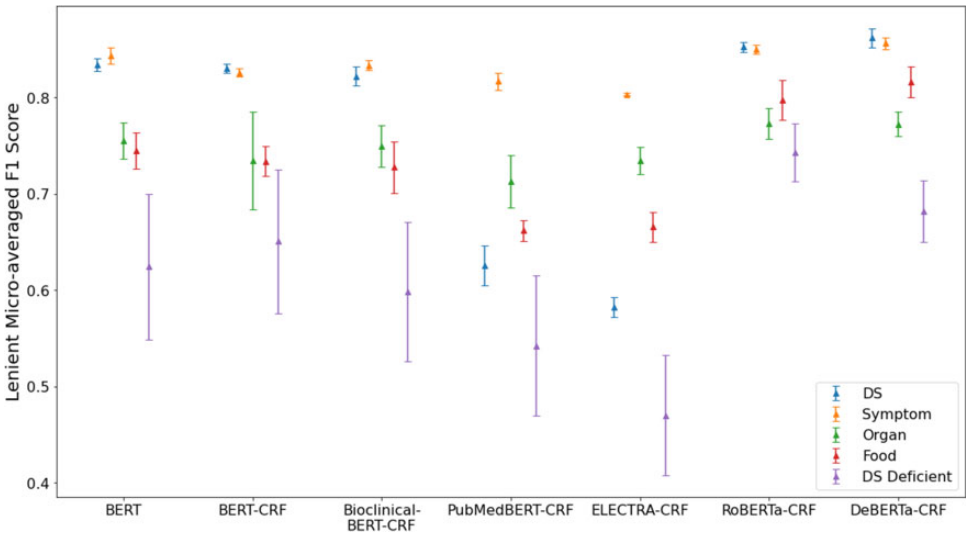

Figure 3. Entity-level F1 scores for concept extraction models.

Table 2. Performance of relation extraction models for DS AEs. The performance of the best model is highlighted in bold.

Model	Precision	Recall	F1
BERT	0.67 ± 0.01	0.79 ± 0.02	0.73 ± 0.01
Bioclinical-BERT	068 ± 0.02	0.79 ± 0.02	0.73 ± 0.02
PubMedBERT	0.68 ± 0.01[a]	0.84 ± 0.02	0.75 ± 0.01
DeBERTa	0.73 ± 0.03	0.82 ± 0.02	0.78 ± 0.02
RoBERTa	**0.74 ± 0.01**	**0.85 ± 0.02**	**0.79 ± 0.01**

DS AE: dietary supplement-related adverse events.

to 2 factors: (1) The DS deficient concepts have the same entity as DS concepts. The "deficiency" is only determined by the context of the word or phrase that corresponds to DS. (2) The DS deficient concepts are scarce in our annotated dataset.

We examined the DS AE relations retrieved from our proposed end-to-end pipeline. Among the 442 227 possible relations between DS and AE symptoms detected by our pipeline, 0.85% were DS AEs, 0.35% DS were deficiency AEs (0.35%), and 3.67% were DS indications. This corroborated that self-reported DS AE signal from Twitter users are scarce among the vast number of tweets.

Vitamins are the most discussed supplements on Twitter, as shown in Tables 4–6. The DS AEs related to vitamins found by our pipeline were mostly related to side effects from taking too much vitamin, as shown in the tweet examples. For nonvitamin DS, 2 frequently mentioned DS AEs caught our attention: (1) fish oil might cause prostate cancer and (2) melatonin might lead to nightmares. The relationship between fish oil usage and prostate cancer risk became popular among media based on the Brasky study, but there is no evidence that fish oil can cause prostate cancer.[46] Melatonin, on the other hand, is frequently used by people with sleeping disorders.[47] While it is still not certain whether the melatonin is the cause of nightmares and crazy dreams. These 2 examples have shown that our proposed model can detect the associations between DS and AE and it could be the basis of further clinical trials or safety test of the supplements.

Error analysis and limitations

Our end-to-end pipeline model was able to achieve almost equal mean F1 scores in identifying DS indications (0.68) and DS AEs (0.61). Our study is similar to SMM4H 2021 Shared Task 1, which extracts drug AE entities and relations from Twitter posts. The end-to-end pipeline performance is on par with one of the best-performing models on the shared task,[36] where the relation extraction model was able to achieve an F1 score of 0.44 and the concept extraction model was able to achieve an F1 score of 0.51. While the decrease in F1 score was expected due to the propagation of concept extraction error through the pipeline, we studied the misclassification details of our pipeline.

The misclassifications could be classified into 3 categories. (1) The pipeline correctly extracted the concepts but labeled the relation wrong. For example, in the tweet "if you are experiencing diarrhea avoid greasy and fried foods caffeine sugary drinks and fruit juices healthy food," caffeine might lead to diarrhea and thus the relation between "caffeine" and "diarrhea" should be a DS AE. The pipeline identified the DS "caffeine" and the symptom "diarrhea" right but labeled the relation as DS indication; (2) the pipeline correctly extracted the concepts but did not give a relation label although a DS AE was in the gold standard. For example, in the tweet "kept on vomiting last night carbonated drinks and caffeine was on the do

Table 3. Performance of the end-to-end DS AEs extraction pipeline

	DeBERTa concept extraction + RoBERTa relation extraction		
Lenient matching	Precision	Recall	F1
DS indications	0.62 ± 0.02	0.76 ± 0.02	0.68 ± 0.01
DS AEs	0.61 ± 0.07	0.62 ± 0.03	0.61 ± 0.04
Overall microaveraged F1 score	0.62 ± 0.01	0.72 ± 0.02	0.67 ± 0.01

DS AE: dietary supplement-related adverse events.

Table 4. Examples of most frequently DS AEs detected by end-to-end deep learning pipeline

DS AE pairs	Frequency	In iDISK?	Tweet examples
Fish oil—prostate cancer	336	No	"Fish oil does not help or prevent heart disease or Alzheimers. It *does* increase prostate cancer. Do not take it"
Vitamin C—kidney stones	165	Yes	"@USER some medications yes. Even prolonged high dose vitamin C causes kidney stones"
Melatonin—dreams	145	No	"Melatonin sure does help me sleep but it also causes some really trippy dreams"
Vitamin D—overdose	114	Yes	"Vitamin D overdose could manifest as persistent vomiting, as was the case for one woman following a knee surgery"
Vitamin B—lung cancer	98	No	"High vitamin B intake may be linked to higher lung Cancer risk in men"
Selenium—prostate cancer	94	No	" Selenium, vitamin E supplements can increase risk of prostate cancer in some men"
Vitamin C—nausea	85	No	"Too much vitamin C or zinc could cause nausea, diarrhea, and stomach cramps. check your dose"
Vitamin C—sick	84	No	"who knew too much vitamin C can make u sick I am upset"
Vitamin D—toxicity	65	No	"Vitamin D2 is a patented drug similar to vitamin D, but is not natural. It ' s been responsible for the majority of toxicity from vitamin D"
Vitamin C—diarrhea	94	Yes	"I would eat this whole bag of oranges, but vitamin C in high doses can induce skin breakouts and diarrhea"

DS AE: dietary supplement-related adverse events.

Table 5. Examples of most frequently DS indication pairs detected by end-to-end deep learning pipeline

DS indication pairs	Frequency	In iDISK?	Tweet examples
Vitamin C—cancer	2611	No	"I agree. Also many studies haven shown high dose treatments of vitamin C are toxic to cancer cells."
Vitamin D—diabetes	2074	Yes	"HEALTH FACT: Vitamin D, omega—3 fish oil and cinnamon can help prevent diabetes."
Vitamin C—skin	1912	Yes	"Ever since I started trying this Vitamin C serum on my face, my skin has been clearing up quite a bit and I am happy ."
Vitamin C—sick	1817	Yes	"Could be entirely in my head but I have been taking vitamin C supplements while I have been sick and I feel a bit better."
Vitamin C—immune system	1681	Yes	"Vitamin C! Need these for my decreased immune system. Sick—ish feeling, please go away! No to cough . . .."
Vitamin D—cancer	1448	Yes	"Vitamin D. High doses. Use that, esp this time of year. Helps with depression, anti—cancer AND mood. It ' s a natural anti—depressant."
Vitamin C—cold	1012	Yes	"So it's Feb . . . I feel a cold coming on. Every f'ing Feb ! ! ! Pumping the green tea & vitamin c!"

not drink list but I still wtf ok self stahp," the caffeine made the user vomit, therefore, the relation between "caffeine" and "vomit" should be DS AE, yet the pipeline did not give any label; (3) the pipeline failed to extract the concept that constitutes a DS AE relation. For example, in the tweet "I have not had crazy outbreaks from this biotin however I have noticed small acne flares very small," biotin could cause an outbreak, therefore, the relation between "biotin" and "outbreak" should be DS AE. However, the pipeline was not able to extract the entity "outbreak," which results in a false negative of DS AE.

Future work

There is still room for improvements for our end-to-end DS AE identification pipeline. We examined our datasets and found that the dataset is highly imbalanced. We calculated the ratio of DS indications to the DS AEs in our annotated dataset, which is 2.88. This implies that DS indications appear almost 3 times frequent as DS AEs. Therefore, we could expand our annotated dataset with more annotated tweets from our original dataset or supplement it with a medication AE dataset such as the SMM4H dataset. We also noticed that DS-related tweets tend to contain misinformation, especially the ones that describe a possible DS indication. For example, "eating more food with vitamin E will benefit your risks from cancer plus it boosts your immune system so it can fight off viruses." To increase the signal-to-noise ratio of our dataset, we could apply veracity analysis on the tweets[46,47] before we start identifying the DS AEs. Furthermore, Twitter users use more colloquial languages to describe their symptom. For example, instead of "vomit" people will more often use "throw up"; another example could be "on fire," which is

Table 6. Examples of most frequently DS AE pairs caused by supplement deficiency detected by end-to-end deep learning pipeline

DS AE pairs	Frequency	In iDISK?	Tweet examples
Lack of Vitamin D—depression	515	No	"Found out yesterday that I am majorly Vitamin D deficient. It explains a lot about my personality and slight depression."
Lack of Vitamin D—dementia	467	No	"Want dementia? Nah, I did not think so. Vitamin D deficiency more than doubles risk of dementia"
Lack of Niacin—Dermatitis	254	No	"Niacin deficiency: signs and symptoms The famous 4 D's: Diarrhoea Dermatitis Dementia Death (if untreated)."
Lack of Niacin—Dementia	254	No	"Niacin deficiency: signs and symptoms The famous 4 D's: Diarrhoea Dermatitis Dementia Death (if untreated)."
Lack of Vitamin D—Fatigue	223	No	"Hey, I know it's summertime for most of you, but we are still computer geeks getting our monitor tans. Quick reminder that low vitamin D (lethargy, fatigue, and poor immune response) is a real, treatable problem that's pretty common in our field."

DS AE: dietary supplement-related adverse events.

a usual expression that describes a "burning sensation." Normalizing these colloquial terms to the biomedical terminologies could help pick out user self-reported DS AEs from noises such as DS advertisement. Finally, we could train a joint-learning model instead of a stepwise pipeline to prevent error propagation.

CONCLUSION

We developed an end-to-end deep learning pipeline to identify DS AEs from tweets. We compared a variety of transformer-based models for concept extraction and relation extraction tasks and assembled an end-to-end pipeline with the best-performing models to detect DS AE signals. DeBERTa-CRF model was the best-performing concept extraction model, achieving a lenient microaveraged F1 score of 0.866. RoBERTa model was the best-performing relation extraction model, achieving a microaveraged F1 score of 0.79. The resulting end-to-end pipeline achieved a lenient microaveraged F1 score of 0.67 and was applied to the entire dataset. We found that our proposed deep learning pipeline not only retrieved DS AEs and DS indications recorded in the current DS knowledge database but also discovered DS AEs and DS indications that were only reported in tweets. Additionally, our end-to-end pipeline also discovered DS deficiency AEs, which were caused by lacking intake of a certain supplement type. The result suggests that Twitter is indeed a complementary source for monitoring DS AEs and our pipeline can detect these signals for further clinical trials or safety research.

FUNDING

This work was supported by the National Institutes of Health's National Center for Complementary & Integrative Health (NCCIH) and the Office of Dietary Supplements (ODS) grant number R01AT009457 (RZ).

AUTHOR CONTRIBUTIONS

RZ and JB conceived the study design. YW carried out the experiments and produced the original draft of the manuscript. All authors contributed to the production of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Rubina Rizvi for her annotation of the tweets.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The annotation guideline and the term lists used to collect the tweets in this article are available in the Supplementary Material online. Other data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

1. Dietary Supplement Use Reaches All Time High. https://www.crnusa.org/newsroom/dietary-supplement-use-reaches-all-time-high Accessed April 13, 2020.
2. Young AL, Bass IS. The dietary supplement health and education act. *Food Drug Law J* 1995; 50 (2): 285–92.
3. FDA 101: Dietary Supplements. U.S. Food and Drug Administration.https://www.fda.gov/consumers/consumer-updates/fda-101-dietary-supplements Accessed April 13, 2020.
4. Geller AI, Shehab N, Weidle NJ, *et al*. Emergency department visits for adverse events related to dietary supplements. *N Engl J Med* 2015; 373 (16): 1531–40.
5. CFSAN Adverse Event Reporting System (CAERS). U.S. Food and Drug Administration https://www.fda.gov/food/compliance-enforcement-food/cfsan-adverse-event-reporting-system-caers Accessed April 13, 2020.
6. Timbo BB, Chirtel SJ, Ihrie J, *et al*. Dietary supplement adverse event report data from the FDA Center for Food Safety and Applied Nutrition Adverse Event Reporting System (CAERS), 2004-2013. *Ann Pharmacother* 2018; 52 (5): 431–8.
7. Ghosh P, Dewanji A. Effect of reporting bias in the analysis of spontaneous reporting data. *Pharm Stat* 2015; 14 (1): 20–5.
8. Toki T, Ono S. Spontaneous reporting on adverse events by consumers in the United States: an analysis of the Food and Drug Administration adverse event reporting system database. *Drugs Real World Outcomes* 2018; 5 (2): 117–28.
9. Golomb BA, McGraw JJ, Evans MA, Dimsdale JE. Physician response to patient reports of adverse drug effects. *Drug Safety* 2007; 30 (8): 669–75.

10. Sloane R, Osanlou O, Lewis D, Bollegala D, Maskell S, Pirmohamed M. Social media and pharmacovigilance: a review of the opportunities and challenges. *Br J Clin Pharmacol* 2015; 80 (4): 910–20.
11. Sarker A, Ginn R, Nikfarjam A, *et al.* Utilizing social media data for pharmacovigilance: a review. *J Biomed Inform* 2015; 54: 202–12.
12. Edo-Osagie O, De La Iglesia B, Lake I, Edeghere O. A scoping review of the use of twitter for public health research. *Comput Biol Med* 2020; 122: 103770.
13. Golder S, Norman G, Loke YK. Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. *Br J Clin Pharmacol* 2015; 80 (4): 878–88.
14. Duh MS, Cremieux P, Audenrode MV, *et al.* Can social media data lead to earlier detection of drug-related adverse events? *Pharmacoepidemiol Drug Safety* 2016; 25 (12): 1425–33.
15. Eberts M, Ulges A. Span-based joint entity and relation extraction with transformer pre-training. In: ECAI 2020; 2020 (325): 2006–13.
16. Uzuner O, South BR, Shen S, *et al.* 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011; 18 (5): 552–6.
17. Kelly L, Goeuriot L, Suominen H, *et al.* Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In: Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, 2014: 172–91.
18. Elhadad N, Pradhan S, Gorman S, *et al.* SemEval-2015 task 14: analysis of clinical text. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015: 303–10.
19. Magge A, Klein A, Miranda-Escalada A, Al-Garadi MA, *et al.* 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 (# SMM4H) Workshop and Shared Task, 2021: 21–32.
20. Wang Y, Adam TJ, Zhang R. Term Coverage of Dietary Supplements Ingredients in Product Labels. *AMIA Annu Symp Proc* 2016; 2016: 2053–61.
21. Rizvi RF, Vasilakes J, Adam TJ, *et al.*, iDISK: the integrated DIetary supplements knowledge base. *J Am Med Inform Assoc* 2020; 27 (4): 539–54.
22. Vig J. A multiscale visualization of attention in the transformer model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019: 37–42.
23. Wang Y, Zhao Y, Zhang J, Bian J, Zhang R. Detecting associations between dietary supplement intake and sentiments within mental disorder tweets. *Health Informatics J* 2020; 26 (2): 803–15.
24. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc* 2015; 22 (3): 671–81.
25. Zhao Y, Guo Y, He X, *et al.* Assessing mental health signals among sexual and gender minorities using Twitter data. *Health Informatics J* 2020; 26 (2): 765–86.
26. Baziotis C, Pelekis N, Doulkeridis C. DataStories at SemEval-2017 task 4: deep LSTM with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017: 747–54.
27. Yuqi S, Jingqi W, Hua X, Kirk R. Enhancing clinical concept extraction with contextual embeddings. *JAMIA* 2019; 26 (11): 1297–304.
28. Mikolov T, Sutskever I, Chen K, *et al.* Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Process Syst* 2013; 26: 3111–9.
29. Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1532–43.
30. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Trans Assoc Comput Ling* 2017; 5: 135–46.
31. Peters ME, Neumann M, Iyyer M, *et al.* Deep contextualized word representations. *In: Proceedings of NAACL-HLT* 2018: 2227–37.
32. Devlin J, Chang M-W, Lee K, Bert TK. Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT* 2019: 4171–86.
33. Wu S, He Y. Enriching pre-trained language model with entity information for relation classification. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management 2019: 2361–4.
34. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020; 36 (4): 1234–40.
35. Ramesh S, Tiwari A, Choubey P, *et al.* BERT based transformers lead the way in extraction of health information from social media. In: Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, 2021: 33–8.
36. Dima GA, Cercel DC, Dascalu M. Transformer-based multi-task learning for adverse effect mention analysis in tweets. In: Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, 2021: 44–51.
37. Zhou T, Li Z, Gan Z, *et al.* Classification, extraction, and normalization: Casia_unisound team at the social media mining for health 2021 shared tasks. In: Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, 2021: 77–82.
38. Alsentzer E, Murphy J, Boag W, *et al.* Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019: 72–8.
39. Gu Y, Tinn R, Cheng H, *et al.* Domain-specific language model pretraining for biomedical natural language processing. *arXiv Preprint arXiv:2007.15779*; 2020.
40. Gururangan S, Marasović A, Swayamdipta S, *et al.* Don't stop pretraining: adapt language models to domains and tasks. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; 2020: 8342–60.
41. Raj Kanakarajan K, Kundumani B, Sankarasubbu M. BioELECTRA: pretrained biomedical text encoder using discriminators. In: Proceedings of the 20th Workshop on Biomedical Language Processing, 2021: 143–54.
42. He P, Liu X, Gao J, Chen W. DeBERTa: decoding-enhanced BERT with disentangled attention. In: *International Conference on Learning Representations;* September 28, 2020.
43. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J Am Med Inform Assoc* 2020; 27 (1): 3–12.
44. Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. *J Am Med Inform Assoc* 2020; 27 (12): 1935–42.
45. Yang X, Yu Z, Guo Y, Bian J, Wu Y. Clinical relation extraction using transformer-based models. *arXiv Preprint arXiv:2107.08957;* 2021.
46. Alexander W. Prostate cancer risk and omega-3 fatty acid intake from fish oil: a closer look at media messages versus research findings. *P T* 2013; 38 (9): 561–4.
47. McGrane IR, Leung JG, St Louis EK, Boeve BF. Melatonin therapy for REM sleep behavior disorder: a critical review of evidence. *Sleep Med* 2015; 16 (1): 19–26.
48. Cheng M, Yin C, Nazarian S, Bogdan P. Deciphering the laws of social network-transcendent COVID-19 misinformation dynamics and implications for combating misinformation phenomena. *Sci Rep* 2021; 11 (1): 1–4.
49. Cheng M, Li Y, Nazarian S, Bogdan P. From rumor to genetic mutation detection with explanations: a GAN approach. *Sci Rep* 2021; 11 (1): 1–4.