Extracting Adverse Drug Events (ADEs) from Medical Forum Posts

Alexander He
alexhe@umich.edu

Abstract

Adverse drug events (ADEs) are an important issue in healthcare, and social media has become a valuable source of information for detecting and monitoring ADEs. In this paper, we focus on identifying ADEs reported by patients in medical forum posts using Named Entity Recognition (NER) techniques. We used the publicly available CSIRO Adverse Drug Event Corpus (CADEC) dataset, which contains 1,250 medical forum posts on patient-reported ADEs. We applied various NER models to identify entities of interest, including drugs, symptoms, and diseases, and evaluated their performance using standard metrics such as precision, recall, and F1 score. Our results showed that NER models based on pre-trained transformers such as XLM-RoBERTa and SpanBERT outperformed the baseline models, achieving micro-average F1 scores of up to 0.73 for the BIOES tagging scheme and 0.77 for the BIO tagging scheme. Fine-tuning embeddings did not improve performance for the Flair-based models but did not negatively impact the transformer-based models. Our study highlights the potential of NER for identifying ADEs from social media and suggests that pre-trained transformer models can be effective for this task. Future work could explore using larger annotated corpora and more advanced NER techniques to improve performance.

Introduction

The goal of my project is to identify Adverse Drug Events (ADEs) reported by patients in medical forum posts. This is a relevant task for similar studies in text data mining, specifically in pharmacovigilance, which refers to "science and activities relating to the detection, assessment, understanding, and prevention of adverse effects or any other medicine-related problem.[1] Before a drug is authorized for use, clinical trials are the single source of data about the safety and efficacy of a drug. After authorization, it would be used in many more patients, for an extended period of time, and with other medicines. In these situations, unforeseen side effects may emerge. Thus, the safety of a drug must be monitored throughout its use in healthcare. One data source for monitoring drug safety is social media, including medical forum posts.

The task of identifying entities and other key information in the text is called Named Entity Recognition, or NER.[2] I will be applying NER to identify ADEs in each text entry. There are multiple ways to conduct NER, including multi-class classification models, conditional random field (CRF) models, and Deep Learning-based models. NLP packages in Python such as spaCy and NLTK have NER functionality, but they are not domain specific and thus cannot be applied to all situations.

Meanwhile, huggingface has a many pre-trained transformers that can be used for NER modeling.

Data

The dataset I found is called the CSIRO Adverse Drug Event Corpus, or CADEC, publicly available at https://data.csiro.au/collection/csiro:10948. As described by the dataset authors, it is a richly annotated corpus of 1,250 medical forum posts on patient reported ADEs, sourced from social media posts that are "largely written in colloquial language and often deviates from formal English grammar and punctuation rules (Karimi et al., 2015)." The authors split the annotation task in two stages. The first stage identified the named entities of interest such as drug names, side effects, diseases, and symptoms. The second stage was terminology association that linked entities to standard terminologies. The dataset includes the tags from both stages (Karimi et al., 2015).

[1] https://www.ema.europa.eu/en/human-regulatory/overview/pharmacovigilance-overview

[2] https://www.geeksforgeeks.org/named-entity-recognition/

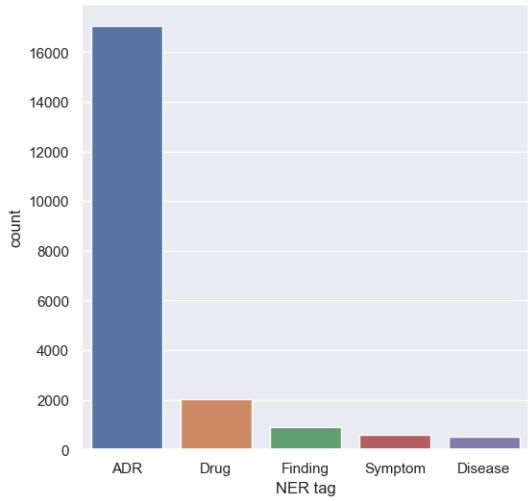

Figure 1: 1st stage tag frequency

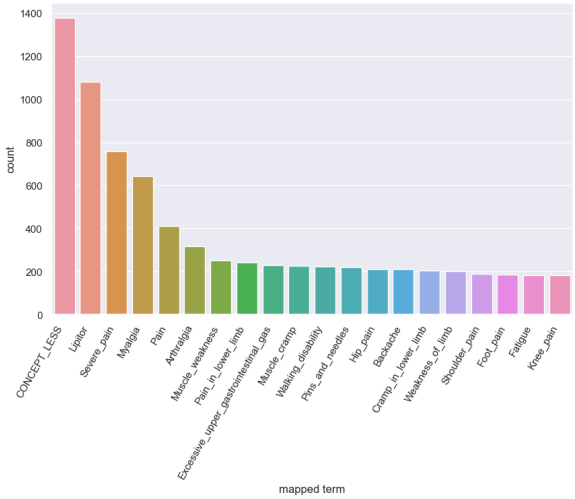

Figure 2: 2nd stage tag frequency (top 20 most frequent)

As part of my analysis of the dataset, I plotted the frequency distributions of the 1st-stage (Fig. 1) and 2nd-stage (Fig. 2) tags.

Related Work

In Polepalli Ramesh et al. (2014), the authors used data from the FDA's Adverse Event Reporting System (FAERS), which is a repository of spontaneously reported ADEs for FDA-approved prescription drugs, including both structured reports and unstructured narratives. Other named entities include drug dosage, frequency and duration, disease symptoms and treatment, and method for administering the drug (the route). They annotated 122 narratives comprising 23,000 tokens as their dataset and created a NER tagger using a supervised machine learning approach. Their best model achieved an F1 score of 73%.

In Li et al. (2018), the authors used data from the Medication, Indication and Adverse Drug Events (MADE) 1.0 challenge. The dataset is a corpus of 1,100 electronic health records (EHRs) of cancer patients. They treated the extraction of ADE-related information as a two-step task: NER and relation extraction. They designed a deep-learning model for extracting ADEs, medications, and indications. They also improved the deep learning model using multi-task learning between the two steps. Their first model achieved an F1 score of 65.9%, and their improved model scored 66.7%, an increase of 0.8%.

In Florez et al. (2018), the authors also used data from the MADE1.0 challenge and used a neural network (or deep learning) approach. Specifically, they used Long Short-term Memory Networks (LSTMs), which can learn long-term dependencies among words in a sentence. For input features, they constructed a comprehensive word representation that concatenates character-level representations, word embeddings, and POS features. The character-level embedding for words was built by a Bi-LSTM network. They achieved a validation F1 score of 71.0% after training on 80% of the data. They also achieved an F1 score of 73.2% when making predictions on the challenge's test dataset after training on 100% of the data.

Methods

The raw dataset is a directory of files. The data preparation required converting these files into a list of objects that the Flair library can use for its Machine Learning models. This was a complex task that required significant debugging to parse the data correctly. The dataset was already clean to a certain extent with few typos, so for tokenization, I treated all words and punctuation as tokens, with the exception of English contractions. For the contractions, I removed the apostrophe and concatenated the result into a single word.

I used the NLP library called Flair,[3] which can use CRF and Bidirectional Long Short-Term Memory (BiLSTM) neural network in tandem for NER. I chose to build NER models with only the 1st-stage tags (ADR, Drug, Finding, Symptom, Disease) due to the class imbalance and limited number of documents in the corpus. The 2nd-stage tags (MedDRA) include 1039 tags, resulting in a much more complex and imbalanced classification prob-

[3] https://github.com/flairNLP/flair/

lem. Building NER models with the 2nd-stage tags would have required a much larger annotated corpus to ensure sufficient representation of each tag, and likely would have resulted in poor performance due to class imbalance. Therefore, I decided to focus on the 1st-stage tags which are most relevant to identifying ADRs, and leave the exploration of the 2nd-stage tagging to future work with a larger annotated corpus.

When designing NER models, a commonly used tagging scheme is the BIO (Beginning, Inside, Outside) scheme. This scheme labels each token in a sequence as either a beginning (B) of an entity, an inside (I) token of an entity, or outside (O) of an entity. For example, in the sentence "Aspirin is a common pain reliever", the drug "Aspirin" would be labeled as "B-Drug", "is" and "a" would each be labeled as "O". The BIOES scheme extends the BIO scheme by adding two additional labels, "E" and "S." The "E" label is used to indicate the last token of a multi-token entity, while the "S" label is used for singleton entities. For example, in the sentence "I have a headache and a fever, I think I have the flu", the word "headache" would be labeled as "B-Symptom", "a" and "fever" would be labeled as "O", and "flu" would be labeled as "S-Disease". The main advantage of the BIOES scheme is that it allows for a more precise labeling of multi-token entities and handling nested entities. However, the BIO scheme is still widely used and can be sufficient for many NER tasks.

I tested a variety of models for comparison. They are all set up similar to the use cases found in the documentation. [4] My base model, indicated by 'flair_original' in the tables, used GloVe[5] and Flair embeddings stacked together as suggested. I also tested removing the RNN component and turning on fine-tuning for the GloVe embeddings, indicated by 'flair_no_rnn' and 'flair_finetune_glove' respectively (Tables 3 and 4). For my other models, I used transformers that are all found on Hugging Face [6] and work with Flair models. I turned on fine-tuning for all these models, similar to few-shot learning. The tested transformers include XLM-RoBERTa, SpanBERT, and Bio_ClinicalBERT.

[4] https://flairnlp.github.io/docs/tutorial-training/how-to-train-sequence-tagger

[5] https://nlp.stanford.edu/projects/glove/

[6] https://huggingface.co/

Evaluation and Results

NER models are evaluated with common metrics such as precision, recall, and F1 score.[7] For baselines, I evaluated random and most-frequent entity prediction models. I did this for annotations from both stages: entities of interest (Table 1) and terminology association (Table 2), which are the 1st and 2nd stages, respectively. The 1st stage has 5 unique tags and the 2nd stage has 1039 unique tags. Thus, it is not surprising that the 1st stage's baselines perform better.

	model	micro_avg_f1
0	ner_random	0.047144
1	ner_mostfreq	0.044356
2	meddra_random	0.010462
3	meddra_mostfreq	0.000335

Table 1: Baseline BIOES models (both stages)

	model	micro_avg_f1
0	ner_random	0.088712
1	ner_mostfreq	0.081700
2	meddra_random	0.010629
3	meddra_mostfreq	0.000502

Table 2: Baseline BIO models (both stages)

Tables 3 and 4 are the model results for the BIOES and BIO models, respectively. You can see that the BIO models perform slightly better for all models. The tags are delineated with less headers, so there are less unique tags to predict overall. Furthermore, for the transformer models, performance seemed to increase with transformer model size.

Fine-tuning the GloVe embeddings significantly reduced performance compared to the base Flair model, where they were not finetuned. Meanwhile, this was not the case for the transformer models.

Removing the RNN layer increased performance. It is possible that the addition of the RNN layer increased the complexity of the model and led to overfitting. By removing the RNN layer, the model may have become simpler and less prone to over-fitting, and also faster), leading to improved performance on the test set.

[7] https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/custom-named-entity-recognition/concepts/evaluation-metrics

	model	micro_avg_f1
0	xlm-roberta-large	0.732026
1	spanbert-large-cased	0.709150
2	Bio_ClinicalBERT	0.694915
3	flair_no_rnn	0.681485
4	flair_original	0.415788
5	flair_finetune_glove	0.083531
6	random_baseline	0.047144
7	mostfreq_baseline	0.044356

Table 3: BIOES models

	model	micro_avg_f1
0	xlm-roberta-large	0.769548
1	spanbert-large-cased	0.758708
2	Bio_ClinicalBERT	0.746886
3	flair_no_rnn	0.694210
4	flair_original	0.470888
5	flair_finetune_glove	0.338586
6	random_baseline	0.088712
7	mostfreq_baseline	0.081700

Table 4: BIO models

Discussion

In this project, I aimed to identify Adverse Drug Events (ADEs) reported by patients in medical forum posts using Named Entity Recognition (NER). I used the Flair library with different pre-trained models to train and evaluate the NER models.

Overall, the NER models achieved better performance than the baselines. The best performing model achieved a micro-average F1 score of 0.732 for the BIOES tagging scheme and 0.770 for the BIO tagging scheme. The results show that pre-trained transformer models like XLM-RoBERTa, SpanBERT, and Bio_ClinicalBERT outperform the approach of using GloVe embeddings with Flair.

To explain the differences between the transformers and Flair and GloVe embedding models, a possible explanation is that GloVe embeddings are trained using a global co-occurrence matrix of words in a corpus. This means that each word's embedding is based on its relationships with all other words in the corpus. As a result, the GloVe embeddings are more general-purpose and may be less specialized for a specific task, such as NER. On the other hand, transformer-based embeddings, such as those used in models like XLM-RoBERTa, are trained on a specific task or set of tasks, such as language modeling or question answering. As a result, these embeddings may already be better suited for a specific NER task out-of-the-box, and fine-tuning may further improve their performance.

Although my models perform quite well compared to those in the related studies, the performance of the models is probably still not satisfactory for end-users of an NLP model in the pharmacovigilance domain. This may be due to the limited size of the annotated corpus and the class imbalance of the data. To improve the performance of the models, more annotated data that covers a wider range of drugs and ADEs is needed.

When comparing the performance of the models to the baselines, the difference in performance was significant. This indicates that the models were able to learn patterns in the data and identify ADEs more accurately than random and most frequent entity prediction models.

The results of this project demonstrate the potential of using NLP techniques to identify ADEs from patient-reported data in medical forums. However, there is still much work to be done to improve the performance of the models and make them useful for end-users in pharmacovigilance. Future work can include exploring different annotation schemes, such as the 2nd-stage tags in the CADEC corpus, and experimenting with different pre-processing techniques to improve the models' ability to handle colloquial language and non-standard grammar. Furthermore, exploring ensemble models and active learning techniques can improve the performance of the models with limited annotated data.

Conclusion

In this project, I applied Named Entity Recognition (NER) to identify Adverse Drug Events (ADEs) reported by patients in medical forum posts. I used the CSIRO Adverse Drug Event Corpus (CADEC), a richly annotated corpus of 1,250 medical forum posts on patient-reported ADEs, sourced from social media (medical forum) posts. My goal was to evaluate different NER models and identify the most effective approach for this task.

After data preparation and experimentation with various NER models, I found that the best-performing model for this task was the XLM-RoBERTa transformer model. The BIO tagging scheme performed slightly better than the BIOES scheme, but the BIOES scheme may be more useful due to its greater identification of structure. Using only the 1st-stage tags (ADR, Drug, Finding,

Symptom, Disease) resulted in a better-performing model compared to using the more complex 2nd-stage tags (MedDRA).

Overall, my project shows the potential for using NER to automatically identify ADEs from medical forum posts. Future work could involve expanding the dataset to include more diverse sources of data and exploring other NLP techniques such as relation extraction. Additionally, there is potential for combining NER with other machine learning techniques to improve performance, such as active learning or transfer learning.

Other Things We Tried

I experimented with other versions of Flair models, such as adding or removing the GloVe or Flair embeddings. Otherwise, most of my time was just trying to get things working, so there wasn't time for me to try anything else.

What You Would Have Done Differently or Next

If I were to start parts of the project over, I would have tried to figure out how to use Google Colab earlier. I did not expect the issues I had with Great Lakes, and switching to Colab proved to be a more efficient solution for training the models. Additionally, I would have explored using more pretrained transformers had I figured out how to use them effectively earlier. Specifically, I would have turned on fine-tuning and used lower starting learning rates as they converge faster with similar results than with higher learning rates. These changes may have allowed me to test a wider range of transformers and potentially improve the performance of the models.

References

Edson Florez, Frederic Precioso, Michel Riveill, and Romaric Pighetti. 2018. Named entity recognition using neural networks for clinical notes. page 7–15.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81.

Fei Li, Weisong Liu, and Hong Yu. 2018. Extraction of information related to adverse drug events from electronic health record notes: Design of an end-to-end model based on deep learning. *JMIR Medical Informatics*, 6:e12159.

Balaji Polepalli Ramesh, Steven M Belknap, Zuofeng Li, Nadya Frid, Dennis P West, and Hong Yu. 2014. Automatically recognizing medication and adverse event information from food and drug administration's adverse event reporting system narratives. *JMIR Medical Informatics*, 2:e10.