Leveraging Natural Language Processing for Automated Drug Advice Extraction and Labeling

Matt Calcaterra

Abstract

Medical documentation can often be dense with complex terms which are difficult for patients to read and interpret. This can cause confusion and decreased education for patients. The application of natural language processing (NLP) techniques and models have the potential to create more easily understandable advice for patient. Two model systems were deployed to handle the tasks of extracting specific advice from medication handouts and labeling the advice with different categories.

A variety of binary relevance models and a transformer based roBERTa model proved to be sufficient for advice labeling. The results further indicated that labeling of advice for patient could be done efficiently with low computational overhead.

The deployment of transformer-based models for advice extraction from raw handout text showed results indicating that they could be viable to patient implementation, however further interactions and refinements will be needed first. BIOE tagging was used to indicate the position of advice in the text, and predict advice position. Condition random field layers added on top of the base transformer models show positive increases in the performance of advice extract, and post-processing treatment of the model predictions could potentially increase this further.

With further refinement of the model processes, the advice extraction and labeling could be combined into a single pipeline for handout extraction and labeling. This would prove to be a pivotal application of NLP for patient implementation to improve health education and outcomes.

1 Introduction

In the realm of healthcare, accurate and efficient extraction of advice from medical texts is crucial for patient safety and informed decision-making. Specifically in the context of patients reading medically dense information, manual extraction of this information is time-consuming and prone to errors. In this project we aim to leverage natural language processing (NLP) techniques to identify advice tags for drug advice handouts and extract advice from drug handouts. The goal was to develop a system capable of identifying and categorizing different types of advice within drug handouts, thereby facilitating easier access to important information for healthcare professionals and patients alike.

This project is specifically targeted towards future implementations of NLP models which would allow for easier digestion of medical texts for patients. Ideally, with enough data, the processes taken in the project can be expanded to other medically dense texts, which once processed will become increasingly easy to digest for patients and improve patient decision-making. However, patients aren't the only ones who should be interested in the results of this project. Physicians and clinical practitioners should also be interested in the results as it will provide a much quicker alternative to manually synthesizing advice from drug handouts.

What this project is not attempting to do is use drug advice handouts to determine specific contraindications, dosages, and adverse effects. Rather, the results of this project aim to show that NLP techniques can be used on medical texts in order to make them more easily digestible by patients, and increase clarity by extracting specific advice from medically dense documents.

The specific tasks which this project aimed to tackle are applying multi-label classification and information extraction. These tasks were addressed through two models, one for multi-label classification of advice text and one for advice extraction. Multi-label classification was deployed through leveraging binary relevance and pre-trained transformers models. Advice extraction was deployed by tagging each word as beginning (B), ending (E), inside (I), or outside (O). This was then passed to a pre-trained transformer to classify advice tokens in order to extract relevant spans of advice.

This differs from previous text extraction of medical information(Sezgin et al., 2023) where the models created in this project are targeted specifically at patients. Previous uses of text extraction in medical data has primary involved named entity recognition (NER) to extract medical data from free-text for the purpose of research or entry into a medical record. Additionally, the models trained in this project are designed to be combined into one coherent pipeline, which differs from previous strictly classification modalities(Prabhakar and Won, 2021).

2 Data

The data for this project was extracted from *A Corpus of Online Drug Usage Guideline Documents Annotated with Type of Advice*[1] (Preum et al., 2018). The Corpus was accessed via download including two forms of data: raw data and annotated data. The annotated data is represented in a tab separated values files, and the raw data includes a directory of sub-directories, where each sub-directory represents a specific drug, and contains a handout.txt file.

2.1 Non-Annotated Data

The non-annotated data consists of individual handout.txt files which contain the full text for specific drug handouts. Each of the handout.txt were accessed and read into a dataset, separating each handout by lines. A short segment of one of the handouts is as follows:

> *Aripiprazole is given by injection into the buttock or upper arm muscle by a health care professional, usually once every month. Some doses of some brands of this medication may also be given once every 6 weeks. Do not rub/massage the injection site after your dose.*

(Preum et al., 2018)

The non-annotated data contains handout text files for 93 different drugs. Each line was numbered upon being read in order to include the line number from the file along with the line instance. It is important to note that the line number in the annotated data does not correspond to these line numbers. The annotated data line numbers were re-assigned during annotation where the lines were determined by periods, where as the line numbers we have assigned on reading were determined by new line characters.

The non-annotated data serves as the input for the information extraction model, with each line from each handout acting as a single input. Additionally, the labels for this data are generated using the annotated advice and BIOE tagging.

2.2 Annotated Data

The annotated data takes the form of a single tab separated value document. This document identifies specific portions of advice from the handouts, and includes annotations of advice tags(Preum et al., 2018). This dataset includes: Drug Name, Drug number, Line number, Advice text, Advice tags, Medication, Food, Activity, Exercise, and disease. This document includes 1005 annotated pieces of advice. Each drug handout has an average of **11.55** annotated advice.

For this project we focused specifically on the advice tags, which are present in the data as ‘AdviceTag1‘, ‘AdviceTag2‘, ‘AdviceTag3‘, and ‘AdviceTag4‘. There are eight unique tags which can be present in the AdviceTag fields which can appear in any of the columns for any drug. The advice tags and distributions are as follows:

Advice Tag	**Count**
Other drugs related	310
Food or beverage related	253
Disease or symptom related	245
Drug administration related	224
Pregnancy related	211
Temporal	182
Activity or lifestyle related	146
Exercise related	40

Table 1: Annotated Data Label Distribution

The annotated data was used as the input for multi-label classification of handout advice. For this task the advice text and advice tags were extracted into a stand alone dataset. Additionally, the advice text was also used to generate the tagging of the non-annotated data for the information extraction model.

[1] https://zenodo.org/records/1173345

3 Related Work

BioWordVec, improving biomedical word embeddings with subword information and MeSH(Zhang et al., 2019)

Offers a demonstration of word embedding in the biomedical world which differs from traditional methods computed at word level. These new methods use word vectors and embedding in order to draw contexts from the text which the word is set in. Our approach differed in that we used supervised learning with a set of predefined labels in order to classify advice in our model.

Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis(Velupillai et al., 2015)

This paper is a literature review and discussion of research from 2008 to 2014. The paper highlights three main areas of improvement for NLP in a clinical setting: 1) more efficient corpus creation, 2) extracting meaning from text, and 3) using NLP techniques and methods for clinical utility. The advancements and improvements found in this paper were used to help generate a model using improved techniques.

Extracting Medical Information From Free-Text and Unstructured Patient-Generated Health Data Using Natural Language Processing Methods: Feasibility Study With Real-world Data(Sezgin et al., 2023)

Sezgin et al. attempt to determine whether an NLP pipeline is feasible for extracting medical information from free-form clinical text. They found that using a combination of NLP techniques was sufficient to create an accurate model for extracting relevant medical information. Similar to this work, we attempted extract specific parts of interest from unstructured text; however, our work differed in that we looked to extract and classify advice through supervised processes.

Text Chunking using Transformation-Based Learning(Ramshaw and Marcus, 1995)

Ramshaw et al. describe their tests of using transformation-based models to employ text chunking. In the paper, they discuss the idea of using BIO tagging to label each token and then use the model to classify texts based on those tags. This system is fundamental to the advice extraction portion of this project. Differing from the Ramshaw's techniques, this project deployed BIOE tagging where the ending of each chunk is also uniquely tagged as E-. This is important in our project since medical texts can include repetitive terms and will help train or model to select distinct spans.

Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention(Prabhakar and Won, 2021)

This work attempts to create a model to classify medical text with the use of deep learning models and attention head. An attention head-based model will be tested for this project in the future, and the findings from this paper regarding the effects of data annotation may prove particularly insightful. While the models deployed in this paper differ from those of this project, insights can be drawn from the deployment of attention heads within medical data.

4 Methodology

Two models were developed and trained in order to extract and label advice from medication handouts. The first model represents an information extraction model which predicts tags of tokens in text as either B-eginning, I-nside, O-utside, E-nd. These were then used to determine spans of tokens which are advice to be extracted from the text. The results of this model can be passed to the second model which is designed and trained for advice labeling. This model uses a transformer-based model to conduct multi-label classification based on the advice tags present in the annotated data.

Prior to the development and training of the two models, the data was pre-processed and baselines were found.

4.1 Data pre-processing

After the loading of the data, a pre-processing step was undergone to create two datasets which were appropriate for the two models: advice extraction and advice classification.

4.1.1 Advice Classification

For the advice classification dataset, the annotated data was reshaped in order to stack the four advice tag columns. Once stacked these tags were encoded with binary labels, with 1 indicating the presence of that tag in any of the four columns and 0 representing an absence in all columns. This processing resulted in a dataset with 11 fields: drug number, line number, and advice text from the original data, and eight fields corresponding to each of the eight advice tags. The data maintained the 1005 rows which were present in the original data.

The distributions of the advice labels in the data can be seen in figure[1]

4.1.2 Advice Extraction

The pre-processing of the data for Advice extraction required the creation of labels for the non-annotated data extracted from the handout files. The original line numbers from the annotated data were reassigned by matching the handout lines and advice text based on drug, then locating the advice text within the raw line data and assigning the corresponding line number. Since the annotated data includes the specific text for advice, rather than the entire line, we were able to use the annotated advice to generate BIOE tagging for the non-annotated data for the purposes of training an advice extraction model.

The advice text was used to tag the corresponding line from the raw data following the BIOE tagging scheme. This represents a modified version of BIO(Ramshaw and Marcus, 1995) (or IOB) tagging where the beginning token of the advice is tagged with B-, the last token of the advice was tagged with E-, the tokens between are tagged with I- indicating inside, and the remaining tokens are tagged with O- for outside. The implementation in the model compares each advice segment to the line text in order to identify the start and end position of the advice within the line. A sequence of ’O’s were generated that matched the length (in words) of the line text. The found start and end positions were then replaced in this list with ’B’ and ’E’ respectively, and all tags between the two were replace with ’I’. This resulted in an array of tags where each tag corresponds to a word in the line text. This process was applied to each line texts in the input data.

4.2 Baselines

Baseline values were created for the two models based on the pre-processed data. The baselines aided in the assessment of the models’ training to determine the magnitude of difference between the models’ predictions and randomly generated predictions.

4.2.1 Advice Extraction Baseline

The baseline for the advice extraction model was created by predicting a random token in the line to label as B- and a random token between B- and the end of the line to label E-. After the beginning and end were determined, those between the two labels were labeled I- and those outside were labeled O-. This procedure generated a random span of tokens from each input. The randomly generated spans were evaluated on two metrics: the accuracy of labels at a token level, and the precision, recall, and F1 at a span level. After evaluating these predictions at the token and span-level we were able to determine our baselines. The baseline results can be seen in Table 2.

Baseline	Metric	Value
Token-level	Accuracy	0.5096
Span-level	Precision	0.2651
Span-level	Recall	0.2698
Span-level	F1	0.2600

Table 2: Advice Extraction Baseline

4.2.2 Advice Labeling Baseline

Initially, the advice labeling model was going to have two generated baselines: one which generated random labels for each input and a second which reported the most common value for each label from the whole dataset. After generating the baselines, the most common value approach was abandoned after determining that the most common value was generating 0 for each label. This is likely due to the sparsity of our labels and distribution of the eight labels. The baseline was set using a system of random predictions, where a binary label was randomly assigned for each of the eight advice label. The predictions based on the input data was then evaluated based on the evaluation criteria described in the evaluation section and can be seen in Table 3.

Baseline	Metric	Value
Random Labeling	Precision	0.2034
Random Labeling	Recall	0.5127
Random Labeling	F1	0.2913

Table 3: Advice Labeling Baseline

4.3 Model Development and Selection

Two model systems were developed to address the two main components: Advice extraction and advice labeling. In order to have the highest rate of success and efficiency, the model for each will be evaluated and selected separately.

4.3.1 Advice Labeling Model Design and Training

The advice extraction task was tested using multiple models, which were compared after evaluation. The two modalities of models were Binary Relevance and Transformer-based multiclassification. The data was split into train test split using a ratio of 0.8/0.2.

For the binary relevance models, the advice text from the annotated dataset was encoded using term frequency-inverse document frequency (TF-IDF)[2]. This allowed for the balancing of tokens importance based on their frequency in the corpus. The encoded training data was looped through by label and the encoded advice and single label were used to fit a model on the training data and predict on the encoded test data. This was repeated for all eight labels until each label was trained and predicted individually. After generating test predictions for all eight labels, the predictions were evaluated using the metrics discussed later. This process was done for three different classifiers: LogisticRegression[3], SVC[4], KNearestNeighbors[5].

The second modality of model developed used a transformer-based model to prediction all eight labels with dependency. The model selected for use was roBERTa[6] due to its high performance in multi-label classification. The initial train test split was used to further split the train data into train and val datasets using a 0.7/0.3 split. The train data was used to train the model, and the val data as the evaluation data for the model trainer. The roBERTa model and tokenizer were loaded using huggingface transformer tools, then the train data was tokenized and encoded using the loaded tokenizer. The tokenization was setup to pad and truncate the data the data based on the max-length, and deployed through batching of the dataset object. After tokenization, the training arguments were established and a trainer object was initialized using the encoded training dataset, encoded evaluation dataset, pre-trained model, and tokenizer. The trainer object was then used to fine-tune the pre-trained model for the down-stream advice classification task. Following training, the withheld text dataset was tokenized and encoded, and predictions were generated for evaluation.

The results of these models were used to determine which modality would be the most effective.

4.3.2 Advice Extraction Model Design and Training

The model developed for advice extraction was a modified version of roBERTa transformer for token classification. Similar to the transformer model in the advice labeling task, the train dataset was split into train and evaluation datasets using a 0.7/0.3 split. The huggingface transformer tool was used to load the tokenizer for the pre-trained roBERTa model. A tokenizing and pre-processing function was created to handle the data encoding. Due to the way the data is labeled (one tag per word), it is necessary to realign the labels with the new tokens since more than one token can be created per word. The pre-processing function tokenizes each word of the input text individually, then, if more than one token is created for that word, the label is kept for the first token, then any remaining subtokens are assigned the label -100. Tokens with a label of -100 were ignored by the loss function in the trainer. The pre-processing function was then mapped to the datasets to encode the data and adjust the labels to match.

A custom transformer class was built that extends the RobertaForTokenClassification model. A Conditional Random Fields (CRF) [7] (Wallach et al., 2004) layer was added on top of the base model class. The CRF layer was added in order to gear the model towards predicting spans of text over individual labels by correcting for relationship between tokens ('I' should never be followed by 'O'). After establishment, a custom trainer object was created that extended the base huggingface trainer. The cross-entropy loss function was adjusted to readjust the weights to prioritize the prediction of the 'B' and 'E' tags, over the 'I' and 'O' tags. This was done to correct for the imbalance of labels in the data.

The default config for the pre-trained model was loaded using the huggingface config tool, and the config and pre-trained model path were passed the

[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[6] https://huggingface.co/FacebookAI/roberta-base

[7] https://pytorch-crf.readthedocs.io/en/stable/

CRF modified model class to initialize the custom model object. The training arguments were then set and the custom trainer object initialized using the encoded training dataset, encoded evaluation dataset, custom model and tokenizer.

The modified trainer was then trained on the training data, and, following training, was used to create predictions on the withheld encoded test dataset. The predictions were then evaluated at the span and token level. A post-processing function was applied to the results which filled all values between the first predicted ’B’ and ’E’ with ’I’. These post-processed results were also evaluated.

This process was repeated for the base roBERTa model with no added CRF layer. The custom trainer was still used to incorporate the custom label weights.

4.4 Evaluation

Due to the nature of the two tasks, each modality will require its own form of evaluation, which will become specifically relevant to ensure that the models are able to function independently from each other. This allows for fine scale adjustment of each model independently.

4.4.1 Advice Extraction Evaluation

The advice extraction model has two separate metrics. The first is evaluation at the token-level, which assesses the model’s performance at predicting the correct BIOE tag for each token in the input. The evaluation compares the predicted tags for the data to the tags generated from the advice text during pre-processing. The number of correctly assigned labels will be determined and used to calculate the accuracy of the predictions and score the model. For the transformer results, all token/label pairs where the label is -100 are not included in this calculation. This also discards the padding tokens which would artificially inflate the accuracy.

The advice extraction model was additionally evaluated at the span-level as well. Spans are extracted from the predicted sequences based on the predicted BIOE tags. The extracted spans are then compared to the ground-truth spans in order to calculate the precision, recall, and F1 scores using the macro scoring system.

The double system of evaluation allows for the assessment of the model to determine that the model is not only able to correctly assign most of the BIOE tags, but that the key features of spans (B- or E-) follow indicated patterns to allow accurate span creation.

4.4.2 Advice Labeling Evaluation

The advice labels will be predicted by the model through multi-label classification. The predicted labels will be evaluated by generating precision, recall, and F1-score. It is important to note that due to the sparsity of our labels, and most advice containing a single table, it is important for use to strictly evaluate our models’ precision and F1-score. There will likely be a larger number of false negatives compared to false positives due to the nature of our data.

5 Results

5.1 Advice Labeling Results

All versions of the advice labeling models showed improvement compared to the baseline. Five methods were tested in total: randomized baseline, binary relevance with logistic regression, binary relevance with SVC, binary relevance with K-nearest neighbors, and roBERTa sequence classification. To benchmark the performance of the various models, we compared their precision, recall, and F1 score metrics against a random baseline. The full results can be seen in Table 4

The random baseline achieved a precision of 0.196951, recall of 0.489137, and F1 score of 0.280827, indicating the difficulty of the multi-label classification task and the necessity for trained models to achieve meaningful results.

The binary relevance models using different classifiers demonstrated significant improvements over the random baseline across all metrics. The binary relevance model using SVC exhibited the highest precision at 0.886207, recall at 0.823718, and F1 score at 0.853821, suggesting its effectiveness in accurately predicting advice labels.

The RoBERTa-based model, trained specifically for multi-label sequence classification, also achieved competitive performance metrics. With a precision of 0.873239, recall of 0.794872, and F1 score of 0.832215, the RoBERTa model showed similar capabilities as the SVC model.

While the logistic regression-based model had the highest precision upon evaluation (0.892), in the context of the advice labeling task, it may not be the best model. Since the intended application of the model is for patients to get clearer and easier interpretation of medical text, precision is an important metric to avoid confusion in the patients;

Model	Precision	Recall	F1-score
Random Baseline	0.196951	0.489137	0.280827
Binary Relevance - Logistic Regression	0.891892	0.634615	0.741573
Binary Relevance - Linear SVC	0.886207	0.823718	0.853821
Binary Relevance - K-Nearest Neighbors	0.826367	0.823718	0.825040
roberta-base	0.873239	0.794872	0.832215

Table 4: Advice Labeling Model Results

however, F1 is the best metric for selecting a model since its scored based on both the recall and precision. Based on the F1 score, the SVC model would be the best selection.

Comparatively, the SVC model and the RoBERTa-based approach showed similarly competitive precision and F1 score metrics. Due to the larger computational and size requirements of the roBERTa model, the SVC model may be a better selection unless further training show more promising results for the roBERTa model.

Since all the models performed significantly better than the baseline evaluations, there is plausible evidence to believe that any of the models could be trained to be sufficient at the advice labeling task. While SVC performed the best against the metrics, the correct model would depend on the application and resources available.

5.2 Advice Extraction Results

We began by establishing a baseline performance using random span prediction at both token and span levels. Subsequently, we compared the baseline with four different configurations of the RoBERTa model: RoBERTa with a Conditional Random Field (CRF) layer, RoBERTa with post-processing, RoBERTa with both CRF layer and post-processing, and a base RoBERTa model without additional layers or post-processing. The full results can be seen in Table 5.

At the token-level accuracy metric, the random span baseline achieved a score of 0.505671. Interestingly, both RoBERTa-based models, with the exception of the non-post-processed roBERTa with CRF layer, exhibited worse token-level accuracy. This indicates that the CRF layer increases the accuracy for token level classification, while the post-processing decreases the accuracy regardless of the presence of a CRF layer.

When evaluating span-level performance metrics, namely precision, recall, and F1 score, all RoBERTa models demonstrated notable improvements over the random span baseline. Specifically, the RoBERTa-CRF model achieved a precision of 0.424237, a recall of 0.656760, and an F1 score of 0.491793. Comparatively, the base roBERTa model with CRF layer achieved a precision of 0.419971, a recall of 0.635378, and an F1 score of 0.476158 on the same training parameters. This difference suggests the incorporation of a CRF layer facilitated better sequence labeling, resulting in more accurate identification of prescription-related spans within the text.

However, postprocessing techniques applied to the RoBERTa-CRF model did not yield further improvements in span-level precision, recall, or F1 score. In fact, post-processing led to a decrease in token-level accuracy and a slight decline in span-level F1 score compared to the RoBERTa-CRF model without post-processing.

Notably, the base RoBERTa model without additional layers or post-processing achieved span-level precision, recall, and F1 score comparable to the RoBERTa-CRF model but without the computational overhead associated with CRF layer integration. Since the span-level is more important to the task than the token level accuracy, the reduced computation model may be more applicable. Furthermore, post-processing only slightly increased the recall and F1 performance of the base RoBERTa model.

6 Discussion

The results from the two tasks showed mixed results. While across the board there were improvements compared to the baselines, the advice extraction models still failed to achieve exceedingly high evaluation scores. The scores achieved by the advice labeling models indicate that they would perform at a satisfactory rate in end-user deployment, however the advice extraction models would not.

The advice extraction model actually performed significantly better than the randomly generated

Model	Accuracy	Precision	Recall	F1-score
Random Span Baseline	0.505671	0.284946	0.283297	0.278494
RoBERTa-CRF	0.505671	0.424237	0.656760	0.491793
RoBERTa-CRF - Postprocessed	0.367838	0.411704	0.679263	0.487452
RoBERTa	0.367838	0.419971	0.635378	0.476158
RoBERTa - Postprocessed	0.383428	0.413437	0.690910	0.488762

Table 5: Advice Extraction Results

span baseline. This indicates that the model may have begun to recognize specific themes or relationships between the tokens in the medication handouts. Despite this, the final scores did not indicate that it would be ready for an end-user. The differences in the score are likely due to the model correctly predicting portions of the advice spans which it had picked up throughout the training. This makes sense why it performed better than a randomized baseline, however, the overall low final scores is likely a result of only extract portions of the correct advice accurately.

The actual results corroborate this observation. The random span baseline achieved a token-level accuracy of 0.505671, whereas the best-performing advice extraction model, the RoBERTa-CRF model, achieved the same token-level accuracy but with a significantly higher span-level precision (0.424237), recall (0.656760), and F1 score (0.491793). Despite this improvement over the baseline, the overall performance of the advice extraction models remained sub-optimal.

This is likely due to one or more of the following reasons: lack of large amounts of data, over-fitting to the training data, and/or insufficient hyper-parameter training. While the dataset for advice extraction contained 4283 lines of text, only 909 contained advice text. After creating a train test eval split, this left only 509 examples to train the model on. It is likely that with more data the models would have been able to identify relationship and pattern in the text better, and create more accurate span predictions. This also ties into the over-fitting of the training data. The models during training start at a relatively high loss ($\tilde{1}$) and ended at a low loss ($\tilde{0}.1$). While a decrease in loss is a good sign for learning from the data, this drastic change in loss is indicative of an overfitting to the data. This is further supplemented by better evaluations on the train and eval datasets compared to the withheld test data. This was addressed through weight decay; however further measures could be taken to further improve these results. Finally, conducting more training and evaluation of different hyper-parameters could significantly improve the results of the model. Specifically, testing different label weightings in the loss function could likely have a large effect. When unweighted, the model would almost never predict a beginning or end label on the training data. This is expected as there is only every one of each of these labels, compared to the large number of inside or outside labels. Tweaking these weights through model iterations in order to find the right balance may prove to be beneficial for model improvement.

The final insight drawn from the advice extraction results is that the post-processing used could be improved. The post-processing showed improvement when applied to the base roBERTa model, however, it decreased the performance of the roBERTa with CRF model and did not improve the base roBERTa as much as the CRF layer. This improvement to the base model indicates that the post-processing may be adjustable to improve the results further of the CRF including models.

In contrast, the advice labeling models performed well across the board. All of the tested models showed a significant improvement over the baseline. This was expected as multi-label classification, once trained, tends to perform well in practice. Another element to this improvement over the baseline could be contributed to the lack of advice which had more than one label. Since the baseline was predicting randomly for all the labels, a much larger proportion of the random predicted labels contained more than one category compared to the actual data. A second baseline could be tested which randomly picks one label, and compare these results to the models and current baseline to determine if there is a difference.

The surprising results for the advice labeling were from the binary relevance models. Based on these results, there appears to be little to no dependency between labels. This makes sense based on the initial data, as a large portion of the advice only had a single label. Based on this, it likely makes the most sense to use a binary relevance model in practice, as they would ignore false dependency in the labels and are generally less resource intensive.

In conclusion, while the advice extraction models showed promise, further refinements in data quantity, model architecture, and post-processing techniques are warranted to enhance their performance for end-user deployment. In contrast, the performance of the advice labeling models, particularly binary relevance approaches, highlights their suitability for practical deployment in prescription handout text analysis tasks.

7 Conclusion

Dense medical texts, even those intended for patients, can often be confusing and lead to misinterpretation or poor health results in patients. In order to address this, we have deployed natural language processing techniques to extract and label advice from prescription drug handouts.

roBERTa, a token classification transformer, was deployed to identify and extract specific advice from these handouts. The model performed better than baseline, and with further data, training, and iterations may achieve results satisfactory for deployment with end-users. Additionally, adding a conditional random field layer on top of the base model proved to increase the effectiveness of the models.

A second set of model were trained with the intent of classifying these pieces of advice with advice categories. This involved the development and training of three binary relevance models, as well as a sequence classification roBERTa model. While the extraction model did not perform at a high enough level to create reliable testing data for labeling, the models exceeded on the annotated data, and could likely be deployed with the end-user.

With further development of the extraction model, the two tasks could be combined into a pipeline for full extraction and labeling of medication advice. This could be employed by patients to decrease their cognitive burden and increase their understanding and comprehension of medical advice pertaining to their medications.

Additionally, the application of natural language processing in the medical field continues to grow at an astounding rate. While most research revolves around extracting medical information for clinical use, the techniques deployed should be investigated for application of extraction for patient use.

All the code and data for this project can be found at https://github.com/Mattcalcaterra/630_final

8 Other Things We Tried

The biggest challenge encountered in development of the project was finding a good system for tagging the extraction data and converting it for transformer input. The initial tagging mechanism involved searching for each word from the advice in the line text and assigning the label based on position. This created multiple issues as it would assign the incorrect labels if the words appeared more than once. We also tried to generate the tag labels during tokenization for the transformer model. While this method may prove to be more effective than adjusting preexisting labels during tokenization, we were not able to get the system to work. This was likely due to slight differences in tokenization between the line text and advice text. This could potentially be addressed in the future by better integration of the offset map returned by the tokenizer.

Another system which was attempted for this project was predicting the spans of texting by training a model to predict the start and end index, rather than the tags of individual tokens. Rather than labeling each token of the line with BIOE tagging, the starting and ending index of the advice were extracted. A model was then attempted to be trained to predict this start and end index. When deployed, the model did not perform significantly better than the baseline, even after interactions of training and adjustments. It is likely that given more time and further research that this method could be improved to show promising results, however the BIOE tagging model showed more promising results through our testing.

9 Future Work

Given more time to work on the project, there are more methods we would try in the future.

The first aspect which should be tested is the comparison of different transformer models. In future work, base BERT and distillBERT models could also be trained and compared to the roBERTa models in both the advice extract and advice labeling tasks. Specifically, for the advice labeling task, it is likely that these other transformers may perform just as well as roBERTa, with less size and computational requirements. This is indicated by the performance of the binary relevance models. Additionally, due to the smaller data size, it is possible that the smaller distillBERT model may perform better than roBERTa in the advice extraction task. A final consideration for transformer model testing is to fine-tune a BERT based model which has previously been trained on medical text. Since these models have deeper encodings and understanding of medical text patterns, they may perform better on the advice extraction task.

Secondly, we would like to test an advice labeling model that uses a transformer for binary relevance. Since the SVC model and roBERTa model performed at a similar level, there is potential that doing with binary relevance with the transformer model could yield the best results. The roBERTa model proved to be efficient at predicting the labels, however, the results of the binary relevance model also indicated that there was little to no dependency between labels. Dropping the dependency element from the transformer model may end in increased prediction results.

10 Acknowledgements

We would like to give special thanks to David Jurgens and the University of Michigan School of Information for the knowledge and resources which made this project possible.

References

Sunil Kumar Prabhakar and Dong-Ok Won. 2021. Medical text classification using hybrid deep learning models with multihead attention. *Computational intelligence and neuroscience*, 2021.

Sarah Masud Preum, Md. Rizwan Parvez, Kai-Wei Chang, and John A. Stankovic. 2018. A Corpus of Online Drug Usage Guideline Documents Annotated with Type of Advice.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Emre Sezgin, Syed-Amad Hussain, Steve Rust, and Yungui Huang. 2023. Extracting medical information from free-text and unstructured patient-generated health data using natural language processing methods: Feasibility study with real-world data. *JMIR Form Res*, 7:e43014.

Sumithra Velupillai, Danielle Mowery, Brett R South, Maria Kvist, and Hercules Dalianis. 2015. Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of medical informatics*, 24(01):183–193.

Hanna M Wallach et al. 2004. Conditional random fields: An introduction. *University of Pennsylvania CIS Technical Report MS-CIS-04-21*, 24:33–42.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):52.

Appendices

A Example Handout

Patient Educationzolpidem oral
IMPORTANT: HOW TO USE THIS INFORMATION: This is a summary and does NOT have all possible information about this product. This information does not assure that this product is safe, effective, or appropriate for you. This information is not individual medical advice and does not substitute for the advice of your health care professional. Always ask your health care professional for complete information about this product and your specific health needs. ZOLPIDEM EXTENDED-RELEASE - ORAL
(ZOL-pi-dem)
COMMON BRAND NAME(S): Ambien CR
USES: Zolpidem is used to treat sleep problems (insomnia) in adults. It helps you fall asleep faster and stay asleep longer, so you can get a better night's sleep. It may also reduce the number of times you wake up during the night. Zolpidem belongs to a class of drugs called sedative-hypnotics. It acts on your brain to produce a calming effect.
HOW TO USE: Read the Medication Guide provided by your pharmacist before you start taking zolpidem and each time you get a refill. If you have any questions, ask your doctor or pharmacist.
Take this medication by mouth on an empty stomach as directed by your doctor, usually once a night. Since zolpidem works quickly, take it right before you get into bed. Do not take it with or after a meal because it will not work as quickly.
Do not crush or chew extended-release tablets. Doing so can release all of the drug at once, increasing the risk of side effects. Also, do not split the tablets unless they have a score line and your doctor or pharmacist tells you to do so. Swallow the whole or split tablet without crushing or chewing.
Do not take a dose of this drug unless you have time for a full night's sleep of at least 7 to 8 hours. If you have to wake up before that, you may have some memory loss and may have trouble safely doing any activity that requires alertness, such as driving or operating machinery. (See also Precautions section.)
Dosage is based on your gender, age, medical condition, other medications you may be taking, and response to treatment. Do not increase your dose, take it more often, or use it for longer than prescribed. Do not take more than 12.5 milligrams a day. Women are usually prescribed a lower dose because the drug is removed from the body more slowly than in men. Older adults are usually prescribed a lower dose to decrease the risk of side effects.
This medication may cause withdrawal reactions, especially if it has been used regularly for a long time or in high doses. In such cases, withdrawal symptoms (such as nausea, vomiting, flushing, stomach cramps, nervousness, shakiness) may occur if you suddenly stop using this medication. To prevent withdrawal reactions, your doctor may reduce your dose gradually. Consult your doctor or pharmacist for more details, and report any withdrawal reactions right away.
Along with its benefits, this medication may rarely cause abnormal drug-seeking behavior (addiction). This risk may be increased if you have abused alcohol or drugs in the past. Take this medication exactly as prescribed to lessen the risk of addiction. When this medication is used for a long time, it may not work as well. Talk with your doctor if this medication stops working well. Tell your doctor if your condition persists after 7 to 10 days, or if it worsens.
You may have trouble sleeping the first few nights after you stop taking this medication. This is called rebound insomnia and is normal. It will usually go away after 1-2 nights. If this effect continues, contact your doctor.
SIDE EFFECTS: Dizziness may occur. If this effect persists or worsens, tell your doctor or pharmacist promptly. This medication may make you sleepy during the day. Tell your doctor if you have daytime drowsiness. Your dose may need to be adjusted.
Remember that your doctor has prescribed this medication because he or she has judged that the benefit to you is greater than the risk of side effects. Many people using this medication do not have serious side effects.

Tell your doctor right away if any of these unlikely but serious side effects occur: memory loss, mental/mood/behavior changes (such as new/worsening depression, abnormal thoughts, thoughts of suicide, hallucinations, confusion, agitation, aggressive behavior, anxiety).

Rarely, after taking this drug, people have gotten out of bed and driven vehicles while not fully awake ("sleep-driving"). People have also sleepwalked, prepared/eaten food, made phone calls, or had sex while not fully awake. Often, these people do not remember these events. This problem can be dangerous to you or to others. If you find out that you have done any of these activities after taking this medication, tell your doctor right away. Your risk is increased if you use alcohol or other medications that can make you drowsy while taking zolpidem.

A very serious allergic reaction to this drug is rare. However, get medical help right away if you notice any symptoms of a serious allergic reaction, including: rash, itching/swelling (especially of the face/tongue/throat), severe dizziness, trouble breathing. This is not a complete list of possible side effects. If you notice other effects not listed above, contact your doctor or pharmacist.

In the US - Call your doctor for medical advice about side effects. You may report side effects to FDA at 1-800-FDA-1088 or at www.fda.gov/medwatch.

In Canada - Call your doctor for medical advice about side effects. You may report side effects to Health Canada at 1-866-234-2345. PRECAUTIONS: Before taking zolpidem, tell your doctor or pharmacist if you are allergic to it; or if you have any other allergies. This product may contain inactive ingredients, which can cause allergic reactions or other problems. Talk to your pharmacist for more details.

Before using this medication, tell your doctor or pharmacist your medical history, especially of: kidney disease, liver disease, mental/mood problems (such as depression, thoughts of suicide), personal or family history of regular use/abuse of drugs/alcohol/other substances, personal or family history of sleepwalking, lung/breathing problems (such as chronic obstructive pulmonary disease-COPD, sleep apnea), a certain muscle disease (myasthenia gravis).

Do not drive, use machinery, or do any activities that require clear thinking after you take this medication and the next day. You may feel alert, but this medication may continue to affect your thinking, making such activities unsafe. This medication may also increase the risk of falls. You may also experience dizziness or blurred/double vision. Do not drink alcoholic beverages. Children may be more sensitive to the side effects of this drug, especially dizziness and hallucinations.

Older adults may be more sensitive to the side effects of this drug, especially dizziness, confusion, unsteadiness, and excessive drowsiness. These side effects can increase the risk of falling.

Before having surgery, tell your doctor or dentist about all the products you use (including prescription drugs, nonprescription drugs, and herbal products).

During pregnancy, this medication should be used only when clearly needed. Infants born to mothers who have taken sedative-hypnotics near the time of delivery may have undesirable effects such as breathing problems or withdrawal symptoms. Discuss the risks and benefits with your doctor.

A small amount of this medication passes into breast milk. Consult your doctor before breast-feeding.

DRUG INTERACTIONS: Drug interactions may change how your medications work or increase your risk for serious side effects. This document does not contain all possible drug interactions. Keep a list of all the products you use (including prescription/nonprescription drugs and herbal products) and share it with your doctor and pharmacist. Do not start, stop, or change the dosage of any medicines without your doctor’s approval.

A product that may interact with this drug is: sodium oxybate.

Other medications can affect the removal of zolpidem from your body, which may affect how zolpidem works. Examples include rifampin, azole antifungals such as ketoconazole, among others.

The risk of serious side effects (such as slow/shallow breathing, severe drowsiness/dizziness, decreased alertness) may be increased if this medication is used with other products that may also affect breathing or cause drowsiness. Therefore, tell your doctor or pharmacist if you are taking other products such as alcohol, other medicine for sleep or anxiety (such as alprazolam, diazepam, lorazepam), muscle relaxants, and narcotic pain relievers (such as codeine).

Check the labels on all your medicines (such as allergy or cough-and-cold products) because they may contain ingredients that cause drowsiness. Ask your pharmacist about using those products safely.

OVERDOSE: If someone has overdosed and has serious symptoms such as passing out or trouble breathing, call 911. Otherwise, call a poison control center right away. US residents can call their local poison control center at 1-800-222-1222. Canada residents can call a provincial poison control center. Symptoms of overdose may include slowed breathing or a deep sleep from which you cannot be awakened.
NOTES: Do not share this medication with others. It is against the law.
As you get older, your sleep pattern may naturally change and your sleep may be interrupted several times during the night. Consult your doctor or pharmacist for ways to improve your sleep without medication, such as avoiding caffeine and alcohol close to bedtime, avoiding daytime naps, and going to bed at the same time each night.
MISSED DOSE: If you miss a dose, do not take it unless you have time to sleep for 7 to 8 hours afterwards.
STORAGE: Store at room temperature away from light and moisture. Do not store in the bathroom. Keep all medications away from children and pets.
Do not flush medications down the toilet or pour them into a drain unless instructed to do so. Properly discard this product when it is expired or no longer needed. Consult your pharmacist or local waste disposal company.
Information last revised August 2016. Copyright(c) 2016 First Databank, Inc.

B Label Distribution

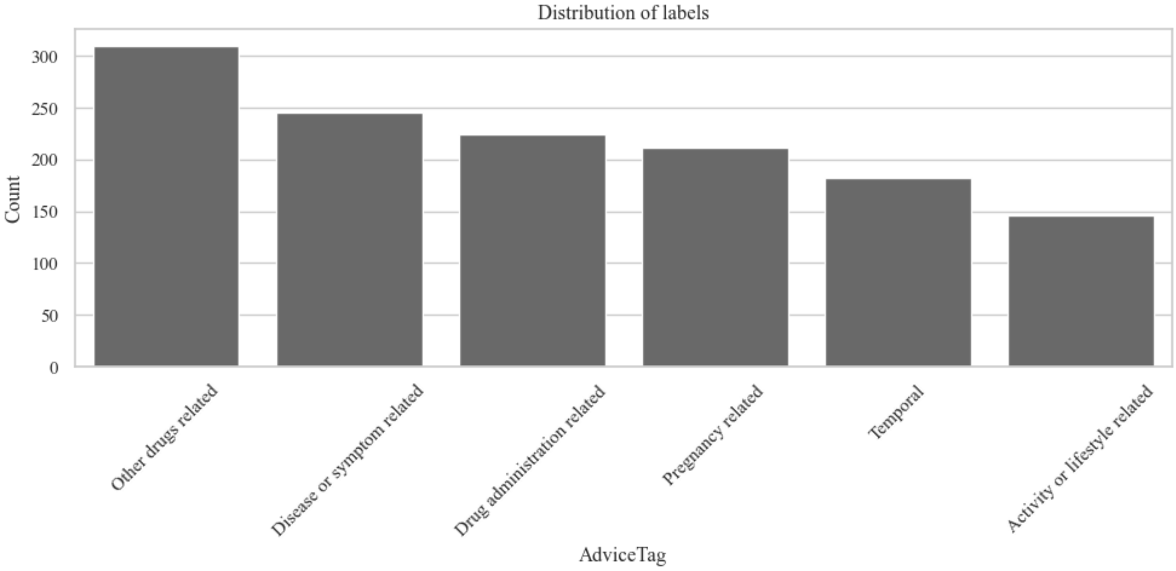

Figure 1: Advice Label Distribution in Annotated Data

C Result Figures

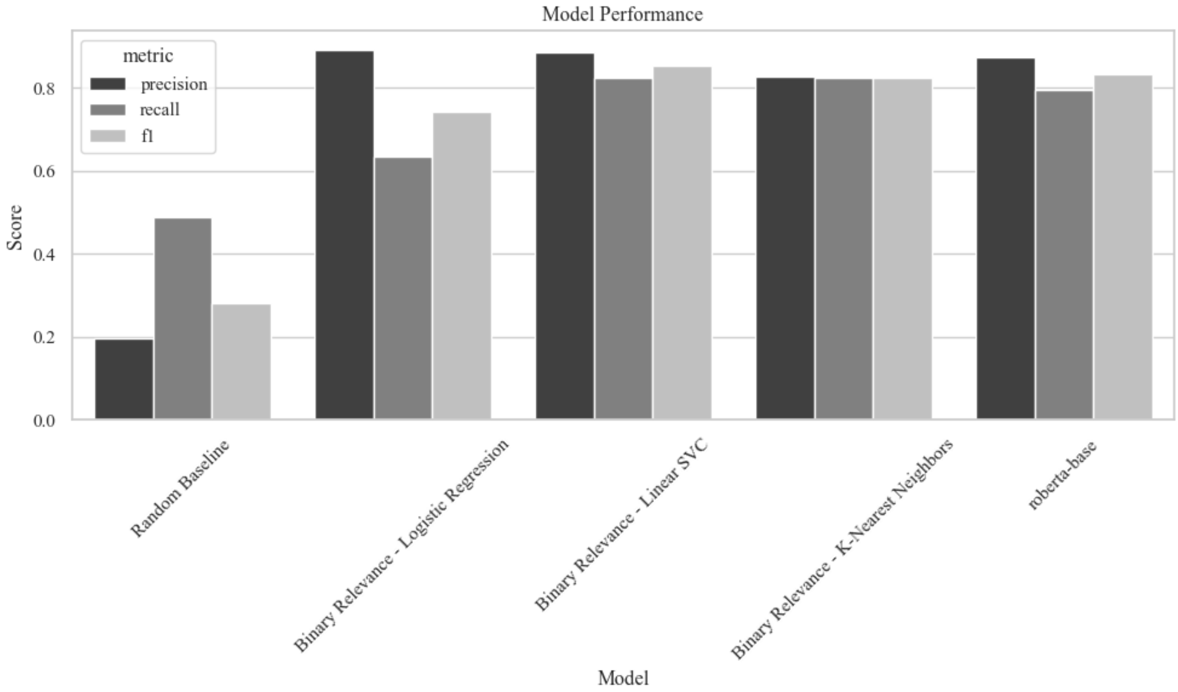

Figure 2: Advice Labeling Model Results

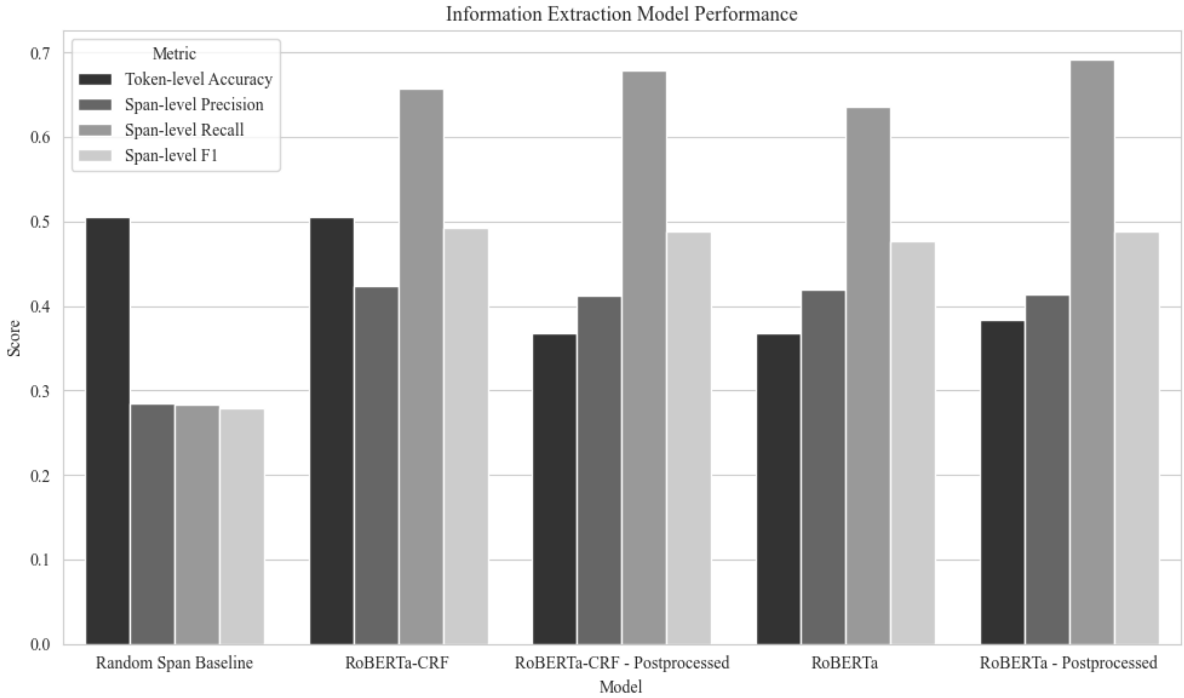

Figure 3: Advice Extraction Results