Document-level Condition-Treatment Relation Extraction on Disease-related Social Media Forums

Sichang Tu
Computer Science
Emory University
Atlanta GA 30322 USA
sichang.tu@emory.edu

Stephen Doogan
Real Life Sciences
Wayne PA 19087 USA
sdoogan@rlsciences.com

Jinho D. Choi
Computer Science
Emory University
Atlanta GA 30322 USA
jinho.choi@emory.edu

Abstract

Social media has become a popular platform where people share information about personal healthcare conditions, diagnostic histories, and medical plans. Analyzing posts on social media depicting such realistic information can help improve quality and clinical decision-making; however, the lack of structured resources in this genre limits us to build robust NLP models for meaningful analysis. This paper presents a new corpus annotating relations among many types of conditions, treatments, and their attributes illustrated in social media posts by patients and caregivers. For experiments, a transformer encoder is pretrained on 1M raw posts and used to train several document-level relation extraction models using our corpus. Our best-performing model achieves the F1 scores of 70.9 and 51.7 for Entity Recognition and Relation Extraction, respectively. These results are encouraging as it is the first neural model extracting complex relations of this kind on social media data.

1 Introduction

There is an increasing number of disease-related posts published online every day. On social media platforms such as Reddit and Twitter, people discuss medical conditions and treatments they use to obtain insights from one another. Capturing medical entities and their relations in these real-world data may significantly benefit tasks such as disease detection (Amin et al., 2020), adverse drug event (O'Connor et al., 2014), and pharmacovigilance (Nikfarjam et al., 2015).

Previous studies have established guidelines and corpora focusing on medical mention, chemical-disease relations, and drug-drug interactions (Uzuner et al., 2011; Patel et al., 2018; Schulz et al., 2020). One limitation of most existing corpora is that their data are collected from well-structured medical text, including electronic health records (EHRs), medical discharges, and clinical notes. Models trained on the corpora of formal medical texts may not perform well on the social media data because social media data are noisy (Baldwin et al., 2013) with poor sentence structures and spelling mistakes. An annotated corpus with carefully designed guidelines is necessary to take full advantage of the large-scale disease-related social media data. However, only a few research works contribute to medical text mining in the social media context (Nikfarjam et al., 2015; Jimeno-Yepes et al., 2015; Basaldella et al., 2020), and no work has directly investigated the condition-treatment relation extraction (RE) on social media data.

To bridge the research gap mentioned above, we develop annotation guidelines and address the automatic extraction of medical entities and condition-treatment relations on social media data (Section 2). Our annotation scheme and the new corpus are illustrated in Section 4. We then experiment with joint models between NER and RE using our corpus (Section 5). Finally, a detailed error analysis of the experiment results is provided in Section 6. The contributions of this paper are as follows:

1. We present annotation guidelines that do not require prior medical knowledge. Unlike many existing medical annotation schemes, our guidelines are not restricted to specific conditions or drugs.
2. We introduce an open-access corpus of 1,150 annotated social media posts in terms of 14 entity types and 2 relation types. To the best of our knowledge, this is the first English condition-treatment RE corpus targeting social media posts.
3. We conduct pilot experiments on automatic entity detection and relation extraction, using a state-of-art document-level joint model. With the pre-trained language model on one million medical social media posts, the best F1 scores for entity detection and relation extraction are 70.9 and 51.7.

Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 218 - 228

2 Related Work

2.1 Medical Datasets

Annotated corpora are essential resources for supervised machine learning. With the advance of NLP in the medical domain, there is increasing research on developing reliable medical corpora for various tasks. For Named Entity Recognition (NER), many datasets are restricted to specific tasks (Uzuner et al., 2008; Uzuner, 2009; Uzuner et al., 2010). For example, in n2c2 datasets[1] (originally known as i2b2), one of their subsets, i2b2 medication dataset (Uzuner et al., 2010) only annotates *Medications* and related entities such as *Dosage*, *Frequency*, and *Duration* in discharge summaries. Moreover, the sources of most datasets are discharge summaries, clinical reports, electronic healthcare records, and biomedical literature.

Very few datasets aim to capture medical entities on social media. Karimi et al. (2015) presented CADEC, the first open-access corpus of medical forum posts. Their corpus comprises 1,321 posts, with annotated entities that are linked to medical terms in controlled vocabularies, such as drug names, adverse drug event, disease, and symptoms. However, one limitation of CADEC is that the corpus only covers 12 drugs and their adverse events. Jimeno-Yepes et al. (2015) introduced a corpus of 1300 posts collected on Twitter, with 3 types of entities: *disease*, *pharmacologic substance*, and *symptom*. Furthermore, they experimented with automatic NER and achieved an F1 score ranging from 55% to 66%. Alvaro et al. (2017) collected 2,000 posts from Twitter and PubMed articles by searching 30 drugs. Annotated entities include *drug* in SIDER database, *disease* and *symptom* in the MedDRA ontology. Scepanovic et al. (2020) obtained 1,980 posts from 18 disease-specific subreddits and annotated *symptom/disease* and *drug names*. They further adopted the BiLSTM-CRF model to extract entities and trained a classifier to categorize the Reddit posts on a large scale.

As for relation extraction, even fewer datasets are available. Uzuner et al. (2011) published the i2b2 clinical relation corpus with 871 annotated clinical records. Their corpus captures the relations in terms of the medical problem–treatment, medical problem–test, and medical problem–medical problem. Segura-Bedmar et al. (2013) provided the DDI Corpus, which annotates the drug-drug interaction in 1,017 documents from the DrugBank database and MedLine abstracts. Focusing on radiology reports, Jain et al. (2021) created the RadGraph dataset, which consists of 4 entity types and 4 relation labels. In addition, the authors developed a benchmark model for relation extraction, with a micro F1 score of 82.3/72.9 on two test datasets.

[1] https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

2.2 Medical Text Mining in Social Media

In the past few years, there has been a surge of interest in social media medical text mining, including tasks such as mental illness detection (Jimeno Yepes et al., 2015; Benton et al., 2017; Gkotsis et al., 2017), pharmacovigilance (MacKinlay et al., 2015; Sarker et al., 2016; Correia et al., 2020), and monitoring epidemic (Drinkall et al., 2022). For medical entity extraction in social media, recent studies show that neural network models (Yepes and MacKinlay, 2016; Scepanovic et al., 2020) outperform traditional approaches using conditional random fields or support vector machine.

3 Data

Our data is collected from various social media forums, using the keyword-based method to filter out disease-related posts. The source sites include online support groups, disease forums, message boards, etc. We obtain approximately one million unlabeled social media posts. Table 1 describes the statistics and source site distributions of the data.

4 Annotation Scheme

4.1 Annotation Environment

The annotation platform used for this project is INCEpTION (Klie et al., 2018), a web-based text annotation environment that allows users to create customized annotation layers and import/export documents in various formats. We created one span layer for entities and one relation layer for relations between entities. Each layer is assigned a tagset that controls the possible values for annotation labels (see Section 4.2 and Section 4.3 for details). Figure 3 shows how the post is annotated in INCEpTION. The dataset is exported in the format of WebAnno TSV 3.3 since it supports custom layers. The format captures document properties, including full text, token positions, token offsets, and annotations on custom layers with disambiguation IDs to identify stacked and multi-unit annotations. Appendix A.1 provides a detailed example of exported annotation in TSV format.

Total Posts	1,068,330
Average Word Count	307.9
Source Sites	*Proportion(%)*
dailystrength	9.90
healthboards.com	9.11
mdjunction.com	4.84
cancercompass.com	4.38
netmums.com	4.01
csn.cancer.org	3.91
alzheimers.org.uk	3.59
celiac.com	3.30
psychforums.com	3.05
experienceproject.com	2.87
addforums.com	2.76
forum.childrenwithdiabetes.com	2.68
alzconnected.org	2.56
ehealthforum.com	2.29
inspire.com	2.06
neurotalk.psychcentral.com	1.98
ibsgroup.org	1.79
crohnsforum.com	1.64
diabetes.co.uk	1.55
cancerforums.net	1.53
depresionforums.org	1.49
exchanges.webmd.com	1.32
ourhealth	1.29
diabetesdaily	1.28
reddit.api	1.28
Other (101)	23.5

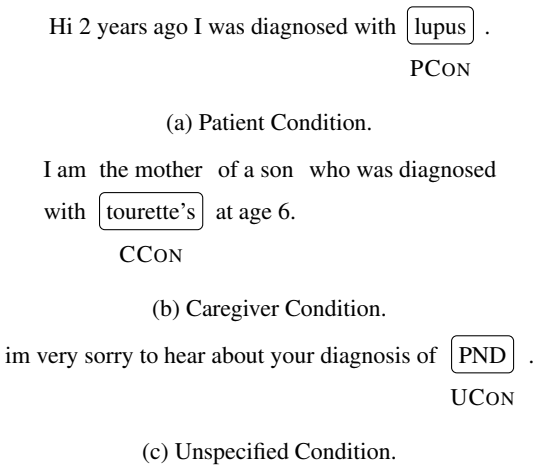
Table 1: Data statistics. **Source sites**: the data distribution of the top 25 sites and the remaining 101 sites.

4.2 Entity Types

The *Entity* layer tagset contains 14 labels in total, which are further divided into 4 subcategories: *Condition*, *Treatment*, *Attribute*, and *Miscellaneous*.

Condition Generally, condition labels capture the disease and any related symptoms, side effects, or impairment caused by the disease or medication. Depending on whom the sufferer is, we annotate the condition as follows:

- PATIENT CONDITION refers to the condition from which the writer of the passage suffers. 'lupus' in Fig 1a is labeled as PATITENT CONDITION since the sufferer is the writer of the post.
- CAREGIVER CONDITION marks the condition affecting someone the writer of the passage cares for (e.g., family members or friends). We annotate 'tourette's' in Fig 1b as CAREGIVER CONDITION, since the patient is the son of the writer.
- UNSPECIFIED CONDITION appears in the context where the sufferer of the condition is unknown or unclear. Another case of UNSPECIFIED CONDITION happens when the condition is assumed or deduced. In Fig 1c, the sufferer is another user in the previous post threads. Hence, 'PND' is labeled as UNSPECIFIED CONDITION.



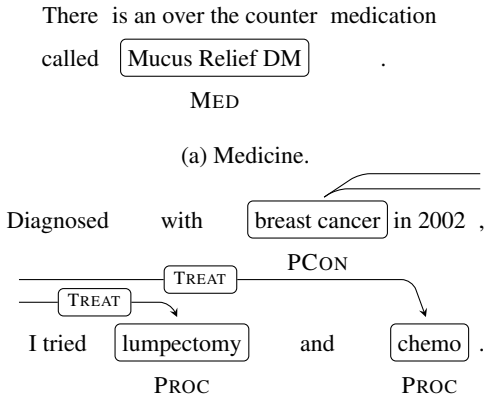
(a) Patient Condition.

(b) Caregiver Condition.

(c) Unspecified Condition.

Figure 1: Examples for *Condition* labels.

Treatment Treatment labels annotate medical treatments (e.g., medicine, surgery, or even counseling) performed to deliver healthcare.



(a) Medicine.

(b) Procedure.

Figure 2: Examples for *Treatment* labels.

- MEDICINE refers to any substance used in treating disease and illness. It could be a drug name, a brand name, or a type of medication. Example is shown in Fig 2a.

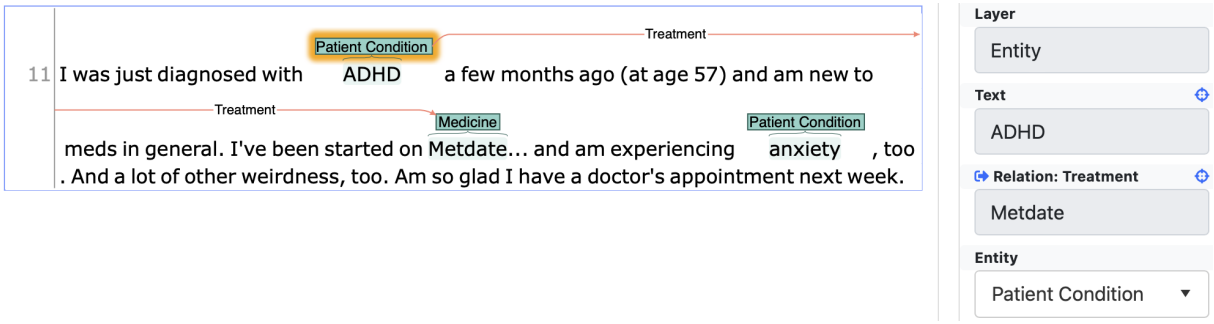
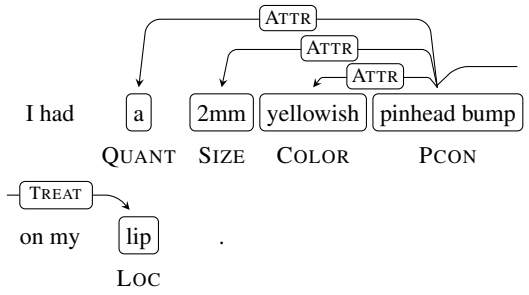


Figure 3: Annotation Interface.

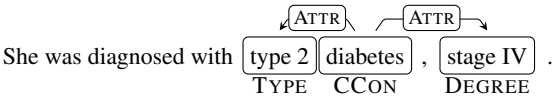
- PROCEDURE marks any medical procedure except for the diagnostic procedure. Common kinds of procedures include surgical procedures (e.g., 'lumpectomy' in Fig 2b) and medical therapy (e.g., 'chemo' in Fig 2b).

Attribute A condition or treatment may have modifiers (usually adjectives or nouns) used attributively to describe them. After carefully examining possible modifier types in the dataset, we conclude 8 attribute labels as follows:

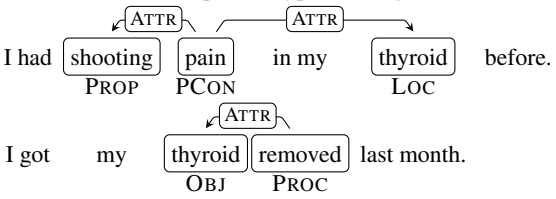
- LOCATION describes where the condition is located or where the treatment happens, such as body parts, anatomical structures, and organs. 'lip' in Fig 4a gives an example for the LOCATION label.
- OBJECT annotates the object to which the treatment is directed. Sometimes it is difficult to distinguish from LOCATION. For example, the 'thyroid' in the second sentence of Fig 4c is labeled as OBJECT since it is the object that was removed. However, the 'thyroid' in the first sentence specifies where the pain occurs and thus is annotated as LOCATION.
- QUANTITY marks the quantity determiner used to specify the condition. It could be concrete numbers (e.g., 'a' in Fig 4a) or quantifiers (e.g., 'several' and 'some').
- COLOR refers to the modifiers that describe the color of the condition. 'yellowish' in Fig 4a gives an example of this label.
- SIZE marks the magnitude and dimension of the condition. It could be linear dimensions (e.g., '2mm' in Fig 4a) or size adjectives (e.g., 'large' and 'small').
- DEGREE shows how severe the condition is, such as disease stages that provides important information on disease development. We label both disease staging (e.g., 'stage IV' in Fig 4b) and adjectives like 'severe' and 'bad' as DEGREE.
- TYPE annotates the specific types of the condition. For instance, 'diabetes' in Fig 4b has three main types, each of which has different symptoms. And the patient is suffering from 'type 2' in the post.
- PROPERTY captures other modifiers that do not fit into the previous attribute labels but provide important properties or characteristics for the condition (e.g., 'shooting' in Fig 4c).



(a) Example for Quantity, Size, Color, and Location.



(b) Example for Type and Degree.



(c) Example for Location, Object, and Property.

Figure 4: Examples for *Attribute* labels.

Note that attribute labels are always attached to corresponding conditions or treatments. Normally, attribute labels would not appear without condition/treatment entities.

Miscellaneous *Miscellaneous* covers entities that do not fit in any of the previous categories, and that may be useful for condition-treatment extraction. Currently, we have one label, PROFILE, under this subcategory. Social media posts on some forums may follow specific conventions, providing additional information after the post content. As shown in Fig 5, the user adds personal information, including username, their relation to the patient, and the patient's medical history to the end of the post. Since it is not a grammatical or complete sentence, we label it as PROFILE separately.

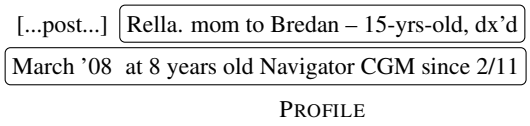


Figure 5: Example for PROFILE.

4.3 Relation Types

Apart from entities, we also annotate directed relations between entities, where applicable. The direction of the relationship is always from the governor to the dependent.

- ATTRIBUTE captures relations between condition/treatment labels and their attribute labels. As shown in Fig 4, all ATTRIBUTE relations go from conditions to attributes. Note that ATTRIBUTE relations are usually intra-sentence relations.
- TREATMENT annotates relations between condition labels and their corresponding treatments. The treatment should be attached to the closest condition with an in-going arc. Fig 2b gives example annotations.

4.4 Corpus Analytics

Since the annotation guidelines we developed require no prior medical knowledge, we recruited undergraduates from Computer Science and Linguistics departments. All annotators went through at least three rounds of annotation training before starting annotation. Initially, 2 annotators were invited and asked to test the guidelines on 6 batches of annotation (10~15 posts per batch). We discussed the issues reported and revised the guidelines accordingly. After this pilot phase, another 2 annotators were recruited to expedite the annotation process. All annotations have been examined and curated by one of the authors.

Table 2 displays the Inter-Annotator Agreement (IAA) scores on the final 3 training batches before the single annotation. Previous study on interrater reliability (Hripcsak and Rothschild, 2005) proves that F1 score is preferable for tasks where the negative case count is unknown or undefined. Our annotation task requires annotators to identify entity boundaries, choose entity labels, and connect relations if applicable. In this case, the annotated entities and relations do not contain any negative cases, which makes traditional metrics such as Cohen's Kappa score inapplicable. Furthermore, calculating the Kappa score on the token level may yield either an unfairly high score if including unannotated tokens or an extremely low score if ignoring unannotated tokens(Brandsen et al., 2020). Hence, F-measure is adopted as the evaluation metric for IAA scores. The F1 score is measured between annotations labeled by annotators and ground-truth annotations we created for the training purpose.

	Round 1 (45)		**Round 2 (50)**		**Round 3 (50)**	
	Ent	Rel	Ent	Rel	Ent	Rel
Annotator 1	42.5	15.9	67.0	44.9	75.5	60.0
Annotator 2	44.5	15.6	69.5	57.9	79.6	76.8
Annotator 3	67.5	55.6	66.2	39.1	78.0	53.5
Annotator 4	73.9	55.9	64.1	44.4	76.5	53.9

Table 2: Inter-Annotator Agreement results measured by F1 score. The number of posts annotated in each round is given in the parenthesis.

On average, we reach an IAA score ~77 for *Entity* and ~60 for *Relation*. Though the IAA scores of *Relations* are lower than *Entity*, note that the relation is correct only if the boundaries and labels of two entities and the relation label are exactly the same. It is noticeable that Annotator 1 and 2 obtained F1 scores ~16 for *Relation* in Round 1. It could be explained by the fact that the guidelines were updated after the two annotators finished the pilot phase, and the agreement scores were measured against the improved ground-truth annotations.

To further analyze the results, we examined annotation disagreements. Disease-related social media data poses certain challenges to the annotation process. First, different from discharge notes or electronic health records, the texts in our dataset use casual language with various expressions to describe the condition/treatment rather than structured formal language with unified medical terminologies. This would lead to the inconsistency of

the entity annotation. Also, the dataset contains considerable long-distance relations, which poses difficulties for annotators to identify the correct governor/dependent entities. Another challenge for annotators is to distinguish between labels such as LOCATION/OBJECT, and PROPERTY/TYPE.

	Count
Total Posts	1,150
Average Word Count	198.53
Entity	9786
Relation	3645

Table 3: Corpus statistics.

Table 3 presents the corpus statistics. We currently have 1,150 annotated posts with 9,786 entities and 3,645 relations. Detailed statistics on specific labels will be provided in Section 5.

5 Experiments

For automatic entity recognition and relation extraction, we adopt the state-of-art joint model for mention detection, coreference resolution, and relation extraction (Xu and Choi, 2022). Focusing on task interactions between mention detection and relation extraction, the model incorporates graph propagation and graph compatibility, which improves decision-making. Since our dataset does not include coreference annotation, the coreference evaluation is not performed in this paper.

Pretraining Though there are existing pretrained language models for the medical domain (Lee et al., 2019; Alsentzer et al., 2019), they are trained on biomedical literature, clinical notes, and discharge summaries. Due to the novelty of the dataset, these language models may not provide good representations for online posts due to the different language styles. To take advantage of pre-trained language models, we continue to train 3 models, namely BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and SpanBERT (Joshi et al., 2020), on 1,068,330 disease-related social media posts.

Preprocessing The joint model requires input documents to be segmented into sentences. Since our annotated dataset is not pre-segmented, one additional preprocessing step is necessary before experimenting with the model. Initially, we utilize ELIT tokenizer (He et al., 2021) to segment posts, followed by remapping all token index, offset and label index. However, the tokenizer fails to process some posts due to the following problems: (1) inappropriate spacing (e.g., lacking space between two sentences in '...vomiting.She...'); (2) unknown characters between entities (e.g., '�' in entity 'Alzeheimer�s'); (3) period after digits and abbreviations (e.g., 'Type 1.' or 'MS.'). Therefore, we create rules to filter out and segment the problematic posts.

Iterative Stratify Split Table 4 shows that the dataset is imbalanced, especially for the entity labels. For instance, PATIENT CONDITION has 2949 instances in the corpus, while COLOR only has 12 instances. As a result, the model generalizability may be hindered, if we randomly split the dataset. To avoid a skewed train/dev/test split, we employ the iterative stratification algorithm designed for multi-label data (Sechidis et al., 2011). Detailed sampling statistics are provided in Table 4.

	Train	Dev	Test	Total
Post	859	118	173	1150
Avg length	198.15	207.71	194.13	198.53
Entity				
PCON	2,189	344	416	2,949
UCON	1,259	173	244	1,676
MED	924	125	165	1,214
CCON	835	94	170	1,099
LOC	565	87	104	756
PROC	492	50	111	653
TYPE	375	60	73	508
DEG	309	44	63	416
PROP	165	18	30	213
OBJ	91	12	16	119
QUANT	72	11	16	99
SIZE	39	4	8	51
PROF	16	2	3	21
COLOR	9	1	2	12
Relation				
ATTR	1,644	246	316	2,206
TREAT	1,088	135	216	1,439

Table 4: Statistics for iterative stratify split.

Results Table 5 gives the results of the joint model on entity recognition and relation extraction tasks using 6 different pre-trained language models as the encoder. It is apparent from this table that by using language models pre-trained on the one million unlabeled data, most models achieve better performance, with an increase ranging from 3.9% to 7.8% for entity and from 3.5% to 5.1% for relation. It is not surprising that SpanBERT-large-med

	Entity			Relation		
	P	R	F1	P	R	F1
BERT-large	55.9	70.1	62.2 (±0.80)	33.1	54.5	41.1 (±0.52)
BERT-large-med	66.8	73.6	70.0 (±0.22)	40.0	**57.6**	47.2 (±1.00)
RoBERTa-large	65.9	73.1	69.3 (±0.53)	42.0	52.9	46.7 (±1.22)
RoBERTa-large-med	65.0	72.3	68.4 (±0.10)	35.9	48.4	41.2 (±1.37)
SpanBERT-large	63.4	71.1	67.0 (±0.57)	45.4	51.3	48.2 (±0.34)
SpanBERT-large-med	**67.5**	**74.9**	**70.9** (± 0.51)	**48.85**	55.0	**51.7** (± 0.66)
Merged Labels						
Condition	68.1	**78.8**	**73.0** (±0.98)	45.4	51.4	48.2 (±1.96)
Treatment	**70.1**	75.2	72.4 (±1.34)	**47.1**	**54.1**	**50.3** (±0.35)

Table 5: Experiment results comparing different pre-trained language models. All scores are the average scores based on 3-5 rounds of experiments. The *med* suffix indicates the model is trained on the one million unlabeled data. The *Merged Labels* section gives results on tagsets with merged labels (e.g., 'Condition' means entity labels under *Condition* subcategory are merged into one tag).

reaches the highest F1 scores for both entity (70.9) and relation (51.7), since it provides an improved prediction on spans and is proved to be promising on span selection tasks.

Besides experimenting with pre-trained language models, we also trained models with various merged tagsets. As in the *Merged Labels* section in Table 5, most merged tagset settings bring an increase on entity F1 scores. However, none of the merged label settings outperforms the original label setting in terms of the relation F1 score.

Postprocessing For this task, higher precision scores are preferable to higher recall scores since less false positive output would benefit the subsequent medical decision-making processes. Hence, we attempt to improve the precision score through postprocessing. First, we adjust the top span ratio for entity extraction, which controls the pruning rate of candidate entities according to their mention scores. Top span ratios ranging from 0.4 to 0.1 are tested, which leads to an average of 1~2 percent increase in precision. Then, we filter out singleton[2] attribute entities with no relation attached, which gives a precision increase of ~2 percent.

6 Error Analysis

Further analysis is conducted based on the model prediction on the test dataset. Table 6 displays the breakdown of results using pre-trained SpanBERT-large model. The best F1 scores are obtained on MEDICINE, CAREGIVER CONDITION, and PATIENT CONDITION since most entities in these labels are likely to be medical terms. The relatively low F1 score of UNSPECIFIED CONDITION is due to mislabeling it as PATIENT CONDITION or CAREGIVER CONDITION. The primary reason for the low performance on OBJECT is that the model is prone to predict anatomical structures as LOCATION. The majority of the entities in PROPERTY are common words, such as 'short' and 'double', which leads to a high false positive rate.

	Count(%)		Results		
	Cor	Spu	P	R	F1
Entity					
MED	86.1	5.5	78.0	80.7	79.3
CCON	84.1	9.4	73.7	76.9	75.3
PCON	83.2	13.5	72.9	75.8	74.4
LOC	81.7	15.4	70.2	77.3	73.6
UCON	77.9	11.5	66.2	68.6	67.4
PROC	73.9	23.4	60.3	70.1	64.8
DEG	61.9	11.1	60.0	57.4	58.6
QUANT	68.8	18.8	57.9	57.9	57.9
PROF	66.7	0	66.7	50.6	57.1
TYPE	57.5	5.5	56.8	53.2	54.9
SIZE	62.5	37.5	45.5	62.5	52.6
COLOR	50.0	0	50.0	50.0	50.0
OBJ	37.5	25.0	40.0	35.3	37.5
PROP	33.3	20.0	26.3	28.6	27.4
Relation					
ATTR	-	-	55.7	58.7	57.1
TREAT	-	-	44.0	48.0	45.9

Table 6: SpanBERT-large-med result breakdown. **Count** shows the proportion of correctly predicted entity count and spurious entity count for each label.

[2]Singleton refers to the single entity without ingoing or outgoing relations attached to other entities.

One crucial problem that is observed from the predicted results is the spurious problem. In other words, the model predicts entities that do not exist in the gold annotation. There is a total of 178 spurious entities produced in 173 posts. The majority of predicted spurious entity types are condition labels (100 spurious entities detected) and treatment labels (35 spurious entities detected). The reasons for this problem are threefold:

1. During the annotation process, we do not annotate singletons such as 'pain' and 'problem' unless they have modifiers that are labeled as attributes (e.g., 'thyroid problem'). The model fails to rule out this kind of singletons.
2. Certain terms such as 'B12' could be treatment for diseases (labeled as MEDICINE in 'B12 supplement') or non-entity (as in 'B12 level'). The model fails to distinguish between these two scenarios.
3. Since most attribute entities we labeled could be non-entity in most times, the model is likely to produce false positive responses. Taking 'severe' as an example, the model may mistakenly label it as DEGREE in 'severe situation', since the model has seen many instances (e.g., 'severe anxiety disorder').

Subsequently, relation extraction also suffers from the spurious problem. Since the relation is generated based on the detected entities, the model would predict relations on spurious entities. Moreover, long-distance relations pose challenges to the relation extraction task. For instance, when more than one condition are labeled in the post, the model is prone to attach the treatment to the closer condition rather than the corresponding one.

7 Conclusion

To facilitate medical text mining in the social media context, we develop an annotation scheme of disease-related posts for the condition-relation extraction. Following the guidelines, we present a reliable corpus[3] with 9,785 entities and 3,645 relations, which is a valuable addition to the limited corpora in this field. Additionally, we experiment with automatic entity recognition and relation extraction, providing a promising model for mining online medical posts. We also conduct a detailed error analysis that may shed light on future work.

The findings of our work suggest potential directions for further studies in this domain. Possible progress could be made by increasing the corpus size since the current corpus is relatively small. Also, the model structure could be designed to solve the spurious problem.

[3] https://github.com/emorynlp/REDSM We distribute the dev and test dataset and part of the training dataset (add up to 50% of the corpus), as requested by the sponsor.

Acknowledgements

We gratefully acknowledge the support of the Real Life Sciences grant. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Real Life Sciences.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Nestor Alvaro, Yusuke Miyao, and Nigel Collier. 2017. Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR Public Health Surveill*, 3(2):e24.

Samina Amin, M. Irfan Uddin, Saima Hassan, Atif Khan, Nidal Nasser, Abdullah Alharbi, and Hashem Alyami. 2020. Recurrent neural networks with tf-idf embedding technique for detection and classification in tweets of dengue disease. *IEEE Access*, 8:131522–131533.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan. Asian Federation of Natural Language Processing.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.

Rion Brattig Correia, Ian B. Wood, Johan Bollen, and Luis M. Rocha. 2020. Mining social media data for biomedical signals and health-related behavior. *Annual Review of Biomedical Data Science*, 3(1):433–458.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Felix Drinkall, Stefan Zohren, and Janet Pierrehumbert. 2022. Forecasting COVID-19 caseloads using unsupervised embedding clusters of social media posts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1471–1484, Seattle, United States. Association for Computational Linguistics.

George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, 7(1):1–11.

Han He, Liyan Xu, and Jinho D. Choi. 2021. Elit: Emory language and information toolkit.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports.

Antonio Jimeno Yepes, Andrew MacKinlay, and Bo Han. 2015. Investigating public health surveillance using Twitter. In *Proceedings of BioNLP 15*, pages 164–170, Beijing, China. Association for Computational Linguistics.

Antonio Jimeno-Yepes, Andrew D. MacKinlay, Bo Han, and Qiang Chen. 2015. Identifying diseases, drugs, and symptoms in twitter. *Studies in health technology and informatics*, 216:643–647.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

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.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

Andrew MacKinlay, Antonio Jimeno Yepes, and Bo Han. 2015. Identification and analysis of medical entity co-occurrences in twitter. In *Proceedings of the ACM Ninth International Workshop on Data and Text Mining in Biomedical Informatics*, DTMBIO '15, page 22, New York, NY, USA. Association for Computing Machinery.

Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Karen O'Connor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, and Graciela Gonzalez. 2014. Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA annual symposium proceedings*, volume 2014, page 924. American Medical Informatics Association.

Pinal Patel, Disha Davey, Vishal Panchal, and Parth Pathak. 2018. Annotation of a large clinical entity corpus. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2033–2042, Brussels, Belgium. Association for Computational Linguistics.

Abeed Sarker, Karen O'connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. 2016. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from twitter. *Drug safety*, 39(3):231–240.

Sanja Scepanovic, Enrique Martin-Lopez, Daniele Quercia, and Khan Baykaner. 2020. Extracting medical entities from social media. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pages 170–181, New York, NY, USA. Association for Computing Machinery.

Sarah Schulz, Jurica Ševa, Samuel Rodriguez, Malte Ostendorff, and Georg Rehm. 2020. Named entities in medical case reports: Corpus and experiments. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4495–4500, Marseille, France. European Language Resources Association.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Özlem Uzuner. 2009. Recognizing Obesity and Comorbidities in Sparse Data. *Journal of the American Medical Informatics Association*, 16(4):561–570.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying Patient Smoking Status from Medical Discharge Records. *Journal of the American Medical Informatics Association*, 15(1):14–24.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Liyan Xu and Jinho Choi. 2022. Modeling task interactions in document-level joint entity and relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5409–5416, Seattle, United States. Association for Computational Linguistics.

Antonio Jimeno Yepes and Andrew MacKinlay. 2016. NER for medical entities in Twitter using sequence to sequence neural networks. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 138–142, Melbourne, Australia.

A Appendix

A.1 Annotation Output

Raw Sentence: I had very bad de realisation when I was first diagnosed with schizoaffective disorder. The doctor came to the house and immediately knew what to do. I had to have a massive dose of tranquillisers over three days. It worked very weel.					
#Text=I had very bad de realisation when I was first diagnosed with schizoaffective disorder .					
1-1	0-1	I	_	_	_
1-2	2-5	had	_	_	_
1-3	6-10	very	_	_	_
1-4	11-14	bad	Degree	Attribute	1-5[1_0]
1-5	15-17	de	Patient Condition[1]	_	_
1-6	18-29	realisation	Patient Condition[1]	_	_
……					
1-13	62-77	schizoaffective	Patient Condition[2]	_	_
1-14	78-86	disorder	Patient Condition[2]	_	_
1-15	86-87	.	_	_	_
#Text=The doctor came to the house and immediately knew what to do .					
2-1	88-91	The	_	_	_
……					
#Text=I had to have a massive dose of tranquillisers over three days .					
……					
3-9	182-196	tranquillisers	Medicine	Treatment	1-13[2_0]
……					
#Text=It worked very well .					
4-1	214-216	It	_	_	_

Figure 6: Exported annotation example after segmentation and remapping (See Section 5). Framed text is the raw text collected from social media forums. Tokens of each sentence have 6 properties: token position (e.g., sentence ID - token ID), token offset, token, entity label, relation label, and disambiguation ID (e.g., governor sentence ID [multi-unit entity ID]).