Knowledge discovery from patient forums: gaining novel medical insights from patient experiences

Dirkson, A.R.

PART III:

EXTRACTING COPING STRATEGIES FOR ADVERSE DRUG EFFECTS

When the dog bites, when the bee stings
When I'm feeling sad
I simply remember my favorite things
And then I don't feel so bad

Rodgers, Hammerstein & Nevin (1981), *The Sound of Music*

8

THE DISCOVERY OF RECOMMENDED COPING MECHANISMS

Edited from: **Anne Dirkson**, Suzan Verberne, Gerard van Oortmerssen, Hans Gelderblom and Wessel Kraaij. How do others cope? Extracting coping strategies for adverse drug events from social media. Submitted.

Patients advise their peers on how to cope with their illness in daily life on online support groups. To date, no efforts have been made to automatically extract recommended coping strategies from online patient discussion groups. We introduce this new task, which poses a number of challenges including complex, long entities, a large long-tailed label space, and cross-document relations. We present the first initial ontology for coping strategies as a starting point for future research on coping strategies, and the first pipeline for extracting coping strategies for side effects. We also compared two possible computational solutions for this novel and highly challenging task; multi-label classification and named entity recognition (NER) with entity linking (EL). We found that coping strategy extraction is difficult and both methods reach limited quality on held out test sets; multi-label classification outperforms NER+EL ($F_1 = 0.220$ vs $F_1 = 0.155$). An inspection of the multi-label classification output revealed that for some of the incorrect predictions, the reference label is close to the predicted label in the ontology (e.g. the predicted label 'juice' instead of the more specific reference label 'grapefruit juice'). Performance increased to $F_1 = 0.498$ when we evaluated at a coarser level of the ontology. We conclude that our pipeline can be used in a semi-automatic setting, in interaction with domain experts to discover coping strategies for side effects from a patient forum. For example, we found that patients recommend ginger tea for nausea and magnesium and potassium supplements for cramps. This can be used as input for patient surveys or clinical studies.

Patients rely heavily on the experiences of other patients for advice on how to cope with their illness in daily life [277]. Specifically, it has been found that patients use online disease-specific forums to gain information from peers [45, 129, 157]. While professionals often approach patients from a primarily medical point of view, patients need to weigh different life values of which 'taking good care of one's body' is but one [49, 56, 238]. Fellow patients are therefore often able to provide more pragmatic and holistic advice to their peers [38].

Adverse Drug Events (ADEs), harmful reactions that result from the intake of medication, are one aspect of their illness that patients need to cope with. ADEs can severely impact the quality of life of patients as well as form a barrier to medication adherence [167]. Although pharmacological management of side effects is sometimes possible, qualitative work indicates that lifestyle and diet can also impact the extent of ADEs, especially for chronic disorders [5].

Previously, qualitative studies have investigated how patients cope with side effects using questionnaires or structured interviews. The most used measurement instrument is the Side Effects Coping Questionnaire (SECope) [148] and the revised version developed by Smedt et al. [279]. It has been employed for the general population [225], patients with HIV [148], and patients with chronic heart failure [279]. The SECope measures general strategies for managing ADE, namely non-adherence, information seeking, social support seeking, and taking medication. The revised version contains two additional strategies: accepting the ADE and requesting other medication from the treating physician.

To date, the only large-scale study into which specific coping strategies patients employ for side effects is an internet survey [156] amongst patients receiving antidepressants. They found that patients employ a variety of methods including changes in lifestyle, diet, and social situations, next to pharmacological management.

Automatic extraction of coping strategies from peer-to-peer resources where patients themselves obtain advice has not been explored. Harvesting coping strategies recommended by patients could provide researchers with new hypotheses and facilitate medical research into which strategies work and why. Some strategies may work to the detriment of medication efficacy. A classic example is the consumption of grapefruit juice which can influence drug metabolism [312]. Our goal is not a fully automated method but a method that produces output that can be assessed and later used by a domain expert.

We focus on coping strategies for adverse drug events specifically. For example[1], in the sentence 'Pickle juice reduces my cramps within just a few minutes' the ADE is cramps and the coping strategy is drinking pickle juice, and in the sentence 'If you feel nauseous, eat ginger' the ADE is nausea and the coping strategy is eating ginger.

The automatic extraction of coping strategies from online patient forums poses four major challenges:

Complex entities The narrative description of coping strategies (e.g. 'take 400mg with breakfast and 400mg with dinner and a big glass of water') results in complex and long entities, which are often not proper nouns. Classic methods for entity extraction are generally not equipped to deal with.

[1] These examples are artificial variants of real sentences in the data to protect patient privacy

No ontology There is at present no ontology to normalize or link the coping strategies to, while aggregation and normalization of coping strategies is vital to be able to provide insight into overall prevalence.

Large and long-tailed label space The large variety of possible coping strategies means that extraction or classification methods will need to be able to deal with a large number of zero-shot cases (i.e. target classes for which there are no examples in the training data) as it is not feasible to collect sufficient data for all target classes.

Cross-document relations Coping strategies are only relevant in relation to a specific ADE and in online discussions these relations may span multiple messages.

An additional complicating factor is that ADE extraction is not trivial. For instance, it is challenging for models to distinguish ADEs from symptoms of the disorder or symptoms resulting from withdrawal (of a medication). The ADE extraction that we employ[2] attains an end-to-end token-level performance of F_1 0.626 and an entity-level performance of 0.716 (Chapter 9).

We address the following research questions:

RQ1 To what extent can coping strategies for ADE be extracted automatically from online patent experiences?

RQ2 How do two approaches to information extraction, namely named entity recognition (NER) with subsequent entity linking and multi-label classification compare on this task end-to-end?

We evaluate our methods on data related to Gastrointestinal Stromal Tumors (GIST), a rare cancer in the digestive system. The Facebook page of the worldwide patient organization GIST Support International (GSI)[3] is the largest online patient community for GIST patients. On the Facebook page, patients share their experiences in discussion threads. The data we work with consists of 124,103 posts in 14,631 threads.

Our main contributions to the medical informatics field are thereby: (1) the novel task of coping strategy extraction, (2) an exploration of extraction and classification methods for its end-to-end resolution and (3) the first ontology for coping strategies. Our code and ontology are publicly available.[4] Unfortunately, our annotated data cannot be shared due to privacy restrictions.

The remainder of the paper is organized as follows: In Section 8.1, we discuss related methodological work. In Section 8.2 and 8.3, we discuss the data sets we use, followed by a detailed description of our methodology. In Section 4.5, we present our results, which are discussed further in Section 8.5.

8.1. RELATED WORK

For the extraction of medical concepts, two broad approaches can be identified. The first approach is Named Entity Recognition (NER) to extract the relevant phrases or

[2] ADE extraction consists of an endr-BERT model and subsequent BioSyn entity linking to SNOMED-CT
[3] `https://www.facebook.com/groups/gistsupport/`
[4] `https://github.com/AnneDirkson/CopingStratExtract`

entities with subsequent entity linking to determine which concept from an ontology is mentioned in the phrase. This approach is widely used for the related task of extraction of ADE from social media messages [193, 266, 335]. State-of-the-art methods for ADE extraction generally rely on domain-specific BERT models [88, 193, 194]. Entity linking of ADE entities is cast as a classification task with all concepts in a medical ontology (e.g., MedDRA or SNOMED-CT) as possible target labels. Because of the large label space, which leads to sparseness in the training data for smaller categories, these methods are designed to be able to deal with zero-shot cases. Similar to coping strategies, the label space for these tasks is both long-tailed and large with over 20.000 labels in MedDRA [194]. Present competitive methods such as BioSyn [291] are often ranking-based and use dense BERT embeddings. The biggest bottleneck at present for end-to-end systems is the extraction step which leads to severe error propagation [194, 335]. Mentions of coping strategies are even longer and more diverse than ADE entities, which makes the problem challenging to be approached as an NER task. The challenge of NER for longer and fuzzy entities has been acknowledged in some recent work, for biomedical concepts [72], human senses [214], motives [332], and emotion causes [182]. We will investigate how well NER with entity linking works for coping strategies using BERT models for NER and BioSyn for entity linking.

The second approach is multi-label classification, which is employed more commonly for tasks such as automatic ICD[5] code assignment [153]. This task is comparable to coping strategy extraction; The label space is also very large and long-tailed (the ICD-9 contains over 15.000 codes and its successor the ICD-10 over 140.000 codes) and multiple labels can be assigned to a single document, i.e., the labels are not mutually exclusive. Although automatic ICD code classification has been explored since the 90s [286], methods have evaluated on the full ICD as opposed to a strict subset of ICD codes only in recent years [212]. While these methods can potentially predict zero-shot cases, they still perform very poorly.

Only a few methods have actually been designed to deal with zero-shot cases to some extent [250, 284]. Rios and Kavuluru [250] extended the CNN-based CAML-DR method of Mullenbach et al. [212] with a graph CNN that makes use of the structure of the label space. Chalkidis et al. [59] find that their model ZAGCNN outperformed transfer learning methods (i.e. BERT and RoBERTA) on few-shot cases and performed comparably on frequent labels. Their results also indicate that exploiting information from label descriptors appears more important than exploiting the label hierarchy for few-shot and zero-shot learning. Song et al. [284] further improve upon the work by Rios and Kavuluru [250] by replacing the CNN with an RNN component. They also propose a latent feature generation framework based on generative adversarial networks [117] to improve the prediction of unseen codes without compromising the prediction of seen codes. Features are generated by exploiting the label structure and label descriptions. As our data does not include label descriptions, these methods are not transferable to the task at hand.

Instead, we opted for a multi-label classification method that does not require label descriptions. We employed a ranking-based (or information retrieval) approach in which labeled data is only used to determine the optimal similarity threshold (i.e., the sentence is

[5] ICD or International Classification of Disease is a terminology for classifying diseases developed by the WHO

Named Entity Recognition (NER)	
CS	781 (2,729 tokens)
– median length CS	3 tokens (mean = 3.55, max = 29)
CS-NEG*	43 (197 tokens)
ADE	2,001 (5,983 tokens)
Negative (O-tag) tokens (included**)	187,355 (95,830)
Posts (included**)	3,715 (1,995)
– median # CS per post	0 (mean = 0.42)
Posts with CS	481
– that also contain an ADE	284 (59%)
Discussions (with CS)	527 (170)
Entity linking (EL)	
CS	824
– with >1 label	59
– with higher order label†	42
# unique concepts	284
% of CSAO in labeled data	0.6%
Posts	481
Multi-label	
CS	824
Posts with CS	481
– median # of labels	1 (max=9)
Negative cases	1514

Table 8.1: Descriptive statistics for Coping Strategy extraction data sets. The multi-label data is converted from the NER and EL data. *Converted to CS for NER **Only a subset of negative examples was included during training †If the concept does not exist in the ontology but the higher order category does

labeled with all labels scoring above this similarity score). Specifically, we used sentence-BERT models to measure the similarity between sentences and target labels. Sentence-BERT models are a class of models introduced by Reimers and Gurevych [247] that are better equipped to handle sentence-level tasks such as multi-label classification. These models employ a pretraining scheme based on Siamese networks.

8.2. DATA

We first detail the data collection and annotation for this novel task in Section 8.2.1 and 8.2.2. The ontology creation is then described in Section 8.2.3. In Section 8.2.4, we describe how we add negative examples to the annotated data.

8.2.1. DATA COLLECTION

In agreement with the GIST Support International Organization, we collected data from their Facebook group. More specifically, we accessed the Facebook official API[6] through

[6] `https://developers.facebook.com/docs/graph-api/`

Relation extraction (RE)	
# of ADE-CS relations	580
– within the same post	397 (68.4%)
– median # of possible ADE per CS	2
– median # co-referents of ADE for which CS is advised	7
# negative cases	1350
median # of annotated posts per CS	6

Table 8.2: Descriptive statistics for the Relation Extraction data between coping strategies (CS) and Adverse Drug Events (ADEs).

a Python script. We got access to the contents of the Facebook group through the account of the group admin. We then collected all posts and comments from the start of the forum. The data ranges from 24 Oct 2009 until 1 Nov 2020 and includes 124,103 posts in 14,631 threads. Our study design was in line with the privacy guidelines of Leiden University and approved by the University privacy officer. The Facebook API did not provide (pseudonymized) usernames in order to protect user privacy. Thus, we were unable to link different posts from the same user within the forum. The collected messages were stored securely, and access was restricted to the involved researchers and annotators. For the labeling of data, we did not use commercial tools but set up private servers that were only accessible to the annotators. In accordance with the GDPR (Article 9.2), we did not obtain consent from each user as the GDPR allows for the use of data from publicly accessible forums with justified cause without individual consent. We are unable to share the data according to the GDPR, because access to the forum has become restricted to members since our data collection (i.e., it is no longer publicly accessible).

8.2.2. DATA ANNOTATION

Named Entity Recognition For annotation, we selected 527 discussions (4,195 posts) based on their likelihood to contain an ADE. We automatically selected the threads that contained at least one drug name according to a match with RxNorm [314]. From these, we selected the threads with the highest percentage of posts in which experiences are shared until our data set included over 4,000 posts. Sharing that someone experienced an ADE falls under this category. In order to estimate which percentage of posts in a thread included patient experiences, we used a previously developed model (Chapter 3).

The data was first annotated by three GIST patients and the first author for the presence of ADEs and coping strategies (CS) for ADE using an annotation guideline.[7] Annotators could indicate with the CS-NEG tag (as opposed to the CS tag) that a coping strategy for an ADE was negative i.e. it entails *not* doing something (e.g. 'avoid salt'). The pair-wise inter-annotator agreement was substantial for ADE (mean κ =0.71) and moderate for CS (mean κ =0.54). The somewhat lower agreement for CS compared to ADE indicates that the CS annotation task is more difficult than the ADE annotation task, but with moderate agreement we still consider the data of sufficient quality to train and

[7] All annotation guidelines are provided at: `https://github.com/AnneDirkson/CopingStratExtract`

Tokens	Pickle	juice	reduces	my	muscle	cramps
NER tags	B-CS	I-CS	O	O	B-ADE	I-ADE
Entity linking	CS04916	CS04916	-	-	-	-

Table 8.3: Example annotation for NER and entity linking

Text	ENTITY_2 (CS)	ENTITY_1 (ADE)	Label*
ENTITY_2 reduces my ENTITY_1 but not my nausea	Pickle juice	muscle cramps	1
ENTITY_2 reduces my muscle cramps but not my ENTITY_1.	Pickle juice	nausea	0

Table 8.4: Example annotation for CS-ADE relation extraction. *1 indicates an CS-ADE relation

evaluate our models on. Data labels were converted to the FuzzyBIO annotation scheme proposed in Chapter 7. We used an online tool Doccano[8] implemented on our own private server for annotation. See Named Entity Recognition in Table 2.3 for details on the annotated data and Table 8.3 for an artificial example of what the annotated data looks like. A more extensive real annotated data fragment is provided in Appendix B (Table B.1).

Normalization The coping strategies were then annotated with concepts from our developed ontology (see Section 8.2.3) by three master students. We switched from Doccano to the annotation tool Inception[9], because Doccano is unable to annotate extracted text spans with concepts from a custom ontology. To switch from Doccano to Inception, we uploaded the earlier NER annotations (in CoNLL-2003 format) from Doccano into Inception. A pilot annotation was used to improve the annotation guideline. All three annotators annotated every post. The inter-annotator agreement was substantial (mean $\kappa = 0.706$) on a token level and moderate (mean $\kappa = 0.475$) on a document (i.e. post) level. Their annotations were curated by the first author. Agreement between at least two of the three annotators was sufficient. The remaining conflicting cases were discussed and resolved. New concepts were added to the ontology where necessary. In 42 cases, the concept was labeled with a higher order concept when the exact concept was not available, e.g., badminton would be labeled with Sport instead of Badminton. If the annotated coping strategy consisted of two strategies (e.g. 'Eat melon and kiwi' or 'Take painkillers and eat well'), the annotators needed to split the strategy to permit labeling. If it was unclear to the annotators what the patient meant, the coping strategy remained unlabeled. This only occurred in 4 cases. See Entity linking in Table 2.3 for details on the annotated data and Table 8.3 for an artificial example. A more extensive real annotated data fragment is provided in Appendix B (Table B.1).

[8] https://github.com/doccano/doccano
[9] https://inception-project.github.io/

ADE-CS relations The annotated coping strategies were coping strategies for a certain ADE. For each CS, three annotators (three different master students) annotated for which ADE the patient recommends the CS. They used the annotation tool Doccano. Annotators were provided with the six messages in the discussion before the post containing the CS. All co-referents of the ADE for which the CS is recommended were annotated. A pilot annotation was used to improve the annotation guideline. Based on an overlapping set of 100 posts, the inter-annotator agreement was measured as the average pair-wise mutual F_1 score of the annotators was 0.757.[10] For every pair-wise calculation, only instances in which at least one of the two annotators found a relation were included. See Table 8.2 for details on the data set and Table 8.4 for an artificial example of what the annotated data looks like. A more extensive real annotated data fragment is provided in Appendix B (Table B.2).

8.2.3. COPING STRATEGY ONTOLOGY

The starting point for our ontology was the experiences of GIST patients we collaborated with and our own experiences with the GIST patient forum. We used these to devise categories of coping strategies patients employ, e.g., edible substances and physical exercise. For each category, we manually selected an appropriate category in one of our source ontologies (e.g., Edible substance (SNOMED-CT 762766007)). We sourced from existing ontologies to allow for interoperability with other ontologies. We chose SNOMED-CT, NCIT and RxNORM as our source ontologies in line with the OHDSI project [222]. We added the PACO Activity Ontology [142] to better represent daily activities and exercise. From the RxNORM ontology we included all Ingredients that are also included in the OMOP vocabulary of the OHDSI project [222]. We used the five hierarchical levels of the ATC (Anatomical Therapeutic Chemical) Classification of the WHO[11] to categorise the RxNORM concepts. The ATC divides medication based on the organ or system on which they act. For normalization, we merged relevant subcategories from different ATC categories into general antibiotics, antiseptics, and antivirals labels, i.e., antiseptics acting on different organs are now grouped.

During annotation, we identified gaps in our ontology. We expanded the ontology with additional categories (e.g., the category 'position of body') and concepts (e.g., 'shampoo' in the existing category 'personal care product' under 'physical object'). These concepts were sourced from the source ontologies if possible. If no appropriate concept was available, we added a concept of our own (e.g. 'split dosage' in the category 'methods of consumption drug' in Table 8.5).

The final ontology contains 48.764 concepts, of which 70.2% from RxNORM, 13.4% from ATC, 9.7% from SNOMED-CT, 6.3% from NCIT, 0.3% from PACO and only 0.1% (64 concepts) were our own additions. The ontology was created using the Python package owlready2. See Table 8.5 for examples and descriptions of the most prominent categories of the Coping Strategy for ADE Ontology (CSAO). We also provide snapshots of the ontology and its hierarchical levels in Table 8.6 and 8.7. The ontology is publicly

[10] The pair-wise F_1 score is preferable to Cohen's kappa for calculating IAA in Named Entity Recognition, as Cohen's kappa needs the number of negative cases which is unknown for NER [41, 138]

[11] `https://www.whocc.no/atc/structure_and_principles/`

Category	Description	Example	# concepts
Adaptation	Includes mental constructs, e.g., attitude and adapting to the circumstances	Positive attitude (SNOMED 225463003)	6
Eating and drinking	Food & drinks, but also frequency and size of meals	Blueberries (SNOMED 227416001)	3,145
Intervention or Procedure	Therapeutic and surgical procedures, alternative therapies and counseling	Thoracentesis (NCIT C15392) Acupuncture Therapy (NCIT C15176)	3,052
Lifestyle	Includes activity, resting, social activities, general dietary recommendations, and clothing strategies	Swimming (PACO 10081)	202
Medication and Supplements	RxNorm medication ingredients categorized by ATC categories	Ondansetron (RxNORM 26225)	40,770
Methods of consumption drug	How and when the medication is consumed	Split dosage (new) After breakfast (SNOMED 7221000175107)	61
Physical object	Various aids, clothing items, and personal care products	Toothpaste (SNOMED 48741003) Single vision glasses (SNOMED 397287009)	1,513
Position of body	Different positions of the body	Sitting (new)	7

Table 8.5: Overview of the major categories in the Coping Strategy for ADE Ontology

available.[12] We consider our ontology – that was initially tailored to GIST – a starting point for more general research into strategies that patients use to cope with side effects.

8.2.4. ADDING NEGATIVE EXAMPLES

Previous work has shown that it is beneficial to include negative examples (i.e., sentences that do not include the item of interest) in the training set for information extraction from medical social media [194]. We found that 481 of the 4,195 posts that were subjected to NER annotation contained coping strategies, thus leaving 3,741 possible negative examples (i.e., sentences that do not contain coping strategies). To reduce the data imbalance, we selected a subset of these negative examples. Specifically, we opted to present the model with difficult negative examples by using forum messages where coping strategies are likely to occur but do not. We accomplished this by selecting the posts that contain an ADE (according to the NER annotation) and the four subsequent messages in

[12] https://github.com/AnneDirkson/CopingStratExtract/blob/main/CSAO.rdf

Eating and drinking	Edible substance	Meat			
		Seafood			
		Dairy food			
		Starchy food	Rice	Brown rice	
				White rice	
				...	
			Bread	Rye bread	
				Tortilla	
				Pita bread	
					White pita bread
					Wholemeal pita bread
				...	
			...		
		...			

Table 8.6: A snapshot of the Edible substance category under Eating and drinking. ... indicate that there are more sub-categories than listed here.

Physical object	Personal care product	Aftershave		
		Baby powder		
		Hair dye		
		Lotion		
		Lip balm		
		Deodorant		
		Mouthwash	Giving analgesic mouthwash	
			Giving antiseptic mouthwash	
			Giving warm saline mouthwash	
		...		

Table 8.7: A snapshot of the Personal care product category under the Physical object section. ... indicate that there are more sub-categories than listed.

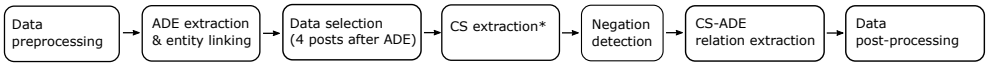


Figure 8.1: Pipeline. ADE: Adverse Drug Effect, CS: Coping Strategy. *Multi-label classification or NER with subsequent entity linking

the discussion. This provided us with 1514 posts (76%) that do not contain CS (see Table 2.3). We included these negative examples in the training set for both NER and multi-label classification.

8.3. METHODS

In Sections 8.3.1 to 8.3.6, we describe the modules of our extraction pipeline for coping strategies shown in Figure 8.1. Although additional components (such as relation extraction, and negation detection) are part of the complete pipeline of extracting coping strategies from online discussions, we define the 'end-to-end' resolution or extraction of coping strategies in this chapter as determining which coping strategies are mentioned in the text.

8.3.1. DATA PREPROCESSING

We preprocessed the data with the pipeline described in Chapter 2. We excluded drug names in the FDA database of drugs[13] from spelling correction to prevent uncommon drug names from being replaced by more common, similar drug names. Removing empty messages and messages in a language other than English left 125,161 messages. Spelling correction corrected 24,834 mistakes. We also normalized drug names to their generic forms using the FDA database.

8.3.2. ADE EXTRACTION AND DATA SELECTION

The extraction of ADE has been described elsewhere (Chapter 9). Adverse drug events were normalized to SNOMED-CT concepts in line with the OHDSI project [222]. Although some previous work has elected to use MedDRA instead of SNOMED, this work focuses predominantly on Twitter data. Annotated datasets for ADE normalization of data that is more comparable to patient forum posts, i.e., Askapatient [151, 353] and Reddit data [20], make use of SNOMED-CT.

For our pipeline, we selected each post that contains an ADE and the subsequent four posts for CS extraction ('Data Selection' in Figure 8.1). Pre-selection of posts that are likely to contain the concept of interest has been shown to aid extraction in social media data with a large signal-to-noise ratio [194]. The window of four subsequent posts was chosen to be relatively wide so as to not miss any coping strategies. The selected posts were not automatically linked to that particular ADE, but purely determined the processing scope for subsequent steps including relation extraction. If an ADE is present in the window of another ADE (e.g., in the second post), its subsequent four posts are also included for CS extraction. The data is deduplicated so any post only occurs once irrespective of the number of ADE within range.

[13] https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files

8.3.3. COPING STRATEGY EXTRACTION

We compared multi-label classification and NER+EL for the end-to-end extraction of coping strategies. These extraction methods are comparable because we know for each sentence which CS concepts it contains.

MULTI-LABEL CLASSIFICATION (MLC)

We used sentence-BERT models [247] for multi-label classification. Sentence-BERT models employ a pretraining method using Siamese networks that results in models more suitable for sentence-level tasks such as measuring semantic similarity. As social media text does not consistently conform to grammatically rules, we choose a pragmatic approach to sentence splitting based on punctuation[14]. We used three different sentence-BERT models [247]: (1) the recommended model for semantic similarity (all-MiniLM-L6-v2) which has been fine-tuned on over 1 billion sentence pairs, (2) a specific natural language inference (NLI[15] model trained on NLI data only and (3) the recommended model for semantic search (msmarco-distilbert-dot-v5) trained on the MSMARCO data set [17]. The MS MARCO data set is a large scale information retrieval corpus based on real user search queries in the Bing search engine and ranked passages for these queries. For this model, the training data consisted of a set of over 500k examples. The full MS MARCO corpus contains over 8 Million examples. The latter model was tuned for dot-product similarity. We also tried the model variant tuned for cosine similarity, but this performed similarly. For the NLI and semantic similarity models, we used the sentences as queries and the labels as retrieval items, whereas for the semantic search model all possible concepts from the ontology (i.e., all possible labels) were used as queries and the sentences as retrieval items because these models are tuned for short queries and longer retrieval documents.

These models were unsupervised and thus training data is not necessary for retrieval. As the models output a similarity (between 0 and 1), we used the training data to determine the optimal threshold (0.1 to 1, steps of 0.1) to select the set of assigned labels. We employed five-fold cross validation in which data are stratified per post.

8

NER WITH ENTITY LINKING

For Named Entity Recognition (NER), we used BERT models, specifically we compared the original BERT model [84] to one trained on English medical social media data (EnDRBERT [303]) and one trained on biomedical texts (PubmedBERT [119]). We used the same five-fold cross-validation as for multi-label classification (60% train, 20% validation, and 20% test per fold). The learning rate (0.01) was optimized on the validation data. Models were trained for 3 or 4 epochs based on validation data. To align experiments with multi-label classification, we trained NER on individual sentences.

We experimented with including ADE as a second entity type during the training of NER models. We expected that identifying ADE may be an easier task than identifying CS and coping strategies for ADE should occur in their vicinity.

We analyzed different possible entity linking methods for the extracted CS phrases. We used the state-of-the-art method for ADE entity linking, BioSyn [291]. We explored

[14] See `https://github.com/AnneDirkson/CopingStratExtract`

[15] Natural language inference is the task of predicting whether one sentence infers the other. An NLI model predicts for a premise whether the hypothesis is true, false or unrelated to the premise.

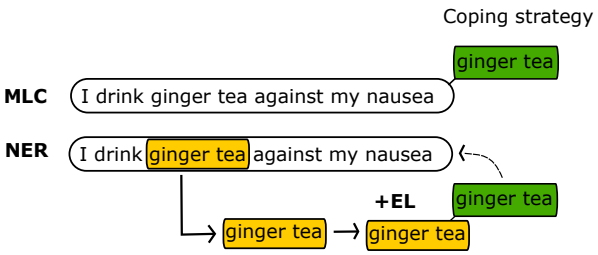


Figure 8.2: Illustration of multi-label classification (MLC) and Named Entity Recognition with Entity Linking (NER + EL). Labels resulting from EL are linked to the original sentence as shown by the dotted line to generate sentence-level results for NER+EL. The sentence-level labels from MLC are then compared with these sentence-level labels from NER+EL.

both BioBERT [174] and SapBERT [187] as base embeddings for this method. SapBERT is a recent pretraining scheme that leverages the UMLS (a biomedical ontology with 4M+ concepts). Liu et al. [187] show that SapBERT pretraining can improve entity linking performance of various BERT-based models with especially large gains for social media data. It also attained a better performance with BioSyn than BioBERT [187]. BioSyn provides a ranking of possible labels present in the phrase. Since CS phrases can have multiple labels, we applied a simple heuristic to allow for multiple labels: The second label is also added if its similarity is closer to the first label than the third label. We attempted to determine a similarity threshold, as we did for the classification approach, but because the similarity metric used in BioSyn is not normalized, this worked poorly.

We compared BioSyn with the best unsupervised multi-labeling classification approach for entity linking. Labels resulting from entity linking were linked to the original sentence to generate sentence-level results for NER+EL. The sentence-level labels from MLC were then compared with these sentence-level labels from NER+EL. Figure 8.2 visualizes this comparison. For these experiments, the same five-fold cross-validation was used.

8.3.4. NEGATION DETECTION

Coping strategies can also entail *not* doing something instead of doing something (e.g. 'I avoid salt'). We found 43 examples during annotation (i.e. labeled CS-NEG) (see Table 8.1). We used a simple heuristic negation method, relying on the Spacy [136] implementation of the Negex algorithm [60]. We used the basic English term set supplemented with additional sixteen preceding and three following heuristics for identifying negation that were manually identified in the data. If one of the heuristics is present, we considered any strategies within the five preceding or subsequent tokens (excluding punctuation) depending on the type of heuristic to be negated. We also determined the dependency relations of strategies. Strategies are negated if they have one of the following dependency relations: (1) negation, (2) no as a determiner or (3) non as a adjectival modifier. We evaluated our heuristic method using entities in the NER that should (CS-NEG) (43 entities) and should not be negated (CS) (781 entities). It attained an F_1 score of 0.810 with a recall of 0.829 and a precision of 0.790.

8.3.5. RELATION EXTRACTION

It is important to determine *which* ADE the coping strategy relates to. We applied a rule-based approach for relation extraction: If there is an ADE mentioned earlier in the message, select the closest one. Otherwise, select the ADE mentioned afterwards within the message. In the annotated data, in 134 of the 365 posts (36.7%) where the ADE is mentioned within the post, another ADE is also mentioned within the post. If there is no ADE in the message itself, select the ADE mentioned closest to the strategy earlier in the discussion within at most preceding four posts.

We evaluated our approach on the annotated data (see Table 8.2). We excluded the 232 cases (29.2%) for which the annotators could not determine which ADE the strategy related to. Manual analysis revealed these were the results of errors in the ADE annotation. Within posts, our rule-based classifier attained an accuracy of 88.4%. For all posts including those with cross-post relations, our classifier attained an accuracy of 84.7%.

8.3.6. DATA POST-PROCESSING

Further data post-processing consisted of three steps. First, we removed strategies that are not connected to any ADE (25.1%) as these are likely to be false positives. We checked a random selection of 50 cases and found that 42 of the 50 were false positives, whereas for the other eight the ADE was missed or not mentioned (e.g., for antidepressants the ADE is implied). Second, we removed labels for which the most important token is already connected to another label with a higher semantic similarity, i.e., a sentence will often be linked to >1 highly similar labels (e.g., 'ground ginger' and 'root ginger' for the token 'ginger' and 'cannabis' and 'cannabis oil' for the token 'marijuana'). We also removed labels for which the most important token is the location of an ADE. The third step was combining multi-label instances; We considered two labels as part of one multi-label instance if the locations of the key tokens are adjacent, they are connected to the same ADE and they have the same negation value. An example is 'high fiber' and 'fruit' for the term 'high fiber fruits'.

8.4. RESULTS

First, we describe our ground truth data in Section 8.4.1. Hereafter, we present the best NER method for extracting spans with coping strategies in Section 8.4.2. We compare the best NER method combined with entity linking with multi-label classification for end-to-end extraction in Section 8.4.3. Section 8.4.4 reports the coping strategies for ADE found in a case study on a patient forum for GIST patients.

8.4.1. DATA DESCRIPTION

As this task is novel, we will describe our ground truth data to explore the challenges this task presents. Table 8.1 describes the annotated data for NER and entity linking. The annotated data contains a total of 824 coping strategies, of which 5.2% were negative strategies meaning they entail not doing something (e.g., not drinking milk). Thus, negation detection will be necessary to differentiate positive from negative strategies. The median length of the annotated coping strategies was relatively short (3 words) but they could be very long (up to 29 words). In fact, 5.4% (52) of the coping strategies contained

	No ADE detection		
	Micro F1	Micro R	Micro P
BERT	0.200 ± 0.157	0.155 ± 0.146	**0.671** ± 0.188
EndrBERT	0.089 ± 0.167	0.089 ± 0.172	0.433± 0.399
PubmedBERT	**0.204** ± 0.170	**0.165** ± 0.160	0.443 ± 0.246
	With ADE detection		
	Micro F1	Micro R	Micro P
BERT	**0.380** ± 0.048	**0.331** ± 0.111	0.522 ± 0.096
EndrBERT	0.251 ± 0.182	0.224 ± 0.205	0.503 ± 0.293
PubmedBERT	0.244 ± 0.119	0.161 ± 0.082	**0.713** ± 0.149

Table 8.8: Token-level evaluation results for NER of coping strategies with or without ADE extraction as a joint task. Our metrics are lenient and ignore prefixes, i.e, it is considered correct when the model predicts the correct entity type for a token irrespective of the B- or I-tag.

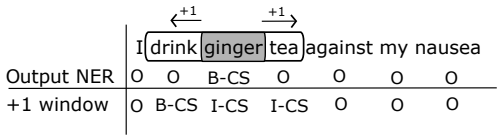


Figure 8.3: Illustration of adding a window of 1 token on both sides of CS mentions identified in NER.

more than 10 words. The data is sparse: Only 11% (481 of 4195) of the posts selected for annotation contained coping strategies. Note that the annotated 527 discussion threads were already preselected to be more likely to contain patient experiences prior to NER annotation so a full patient forum is likely to be more sparse still (See Section 8.2.2).

The ground truth for entity linking demonstrates that not all coping strategies can be captured with a single label from the ontology: 7.2% (59) of the annotated coping strategies were labeled with two labels (e.g. 'cinnamon' and 'chewing gum' for the entity 'cinnamon gum'). Moreover, our ground truth reflects the long-tailed label space. Our labeled 824 coping strategies only cover 284 unique concepts, which equals 0.6% of the ontology.

Table 8.2 describes the ground truth for Relation Extraction between ADEs and coping strategies. On average, there were two different ADEs that the strategy could be linked to within the span of six posts (the post itself and five prior). The ADE for which the CS was advised was mentioned often (an average of 7 times within the span of the post itself and five posts prior). In 31.6% of the cases, the relation was not within the same post but spanned across posts.

8.4.2. NAMED ENTITY RECOGNITION

The first approach to extraction that we evaluated consists of two steps, namely NER and entity linking. Table 8.8 shows the results for the first step of this approach: Named Entity Recognition of coping strategies. We compare models on their micro F_1 score, because it takes into account the label imbalance by aggregating the contributions of all

	Token level evaluation		
	Micro F1	Micro R	Micro P
No window	0.380 ± 0.048	0.331 ± 0.111	**0.522** ± 0.97
+1 on both sides	**0.394** ± 0.018	**0.453** ± 0.108	0.376 ± 0.068
	Entity level evaluation		
	Missed (%)	Correct (%)	Partially correct (%)
No window	**39.1** ± 1.2	27.8 ± 10.9	**33.1** ± 4.1
+1 on both sides	37.2 ± 11.1	**40.0** ± 11.2	22.7 ± 2.5

Table 8.9: Results for adding a window (+1 token) on either side of the extracted CS in NER.

classes and is standard in evaluating multi-label classification tasks. The best performing model was the standard BERT model that was trained to identify both ADE and CS entities ($F_1 = 0.380$). Adding ADE as an additional entity type[16] doubled its performance (+0.180) (See Table 8.8). Without the addition of ADE entities, PubmedBERT, which is trained on biomedical text, outperformed the other models ($F_1 = 0.204$).

Due to the complexity of the CS entities, we explored whether adding an additional token on either side of the identified strategies would benefit performance (See Figure 8.3). Table 8.9 reveals that adding a window of 1 token boosted token-level performance slightly ($F_1 = 0.394$) by increasing recall (+0.122) at a cost to precision (-0.146). On an entity level, the number of entities that are missed entirely was reduced (-1.9 % point), the number of entities that were partially correct was also reduced (-10.4% point), whereas the number of fully correct entities was increased (+12.2% point). We thus included a window of one token on each side for the extracted phrases (i.e., the input for entity linking).

8.4.3. END-TO-END EXTRACTION

Table 8.11 shows the results for end-to-end extraction of coping strategies for both approaches (NER with entity linking and MLC). Although the other multi-label classification models performed very poorly, the best performing method for end-to-end extraction was multi-label classification with the Semantic Similarity sentence-BERT model ($F_1 = 0.220$). With oracle NER (using the manually labeled NER data as input), entity linking using BioSyn based on SapBERT could outperform the classification approach ($F_1 = 0.241$). This higher performance was mainly driven by a higher precision (0.271). Yet, with the addition of NER as an intermediate step the performance dropped below that of multi-label classification. Moreover, multi-label classification outperformed even oracle NER in terms of recall (0.306 compared to 0.283). Macro F_1 scores are computed by averaging the F_1 scores for each class, thus treating all classes equally irrespective of their prevalence. Table 8.11 shows that the macro F_1 scores were far lower than the micro F_1 scores, indicating that across the board the models performed worse on less frequent coping strategies in the annotated data.

As the ontology is hierarchical, we also investigated how far off the predictions of

[16] On a token level, this means adding B-ADE and I-ADE tags

Prediction	Ground truth	Shared higher level
Lip balm	Lotion	Personal care product
Take whole dosage at once	Split dosage	Dosage
Rice	Bread	Starchy food
Therapeutic bed	Assistive bed	Sleeping aid

Table 8.10: Examples of cases where the predicted label and true label are not the same but do fall under the same direct hierarchical category (+1 level)

the best model were by investigating the performance at coarser levels of the concept hierarchy. The results are shown in Table 8.12. The performance was increased to $F_1 = 0.318$ when we considered if the target and predicted labels fell directly under the same direct category in the hierarchy (i.e. '+1 level (strict)') (see Table 8.10 for examples). Also the precision was increased (0.172 to 0.304). The macro F_1 showed a similar increase (from 0.105 to 0.320) which may indicate that it is mostly the infrequent coping strategies that are predicted incorrectly on the detailed level but correctly on the coarser level.

This is rather restrictive measure however, as the target and predicted labels need to fall directly under the same category. There may also be cases where the predicted label is equal to the category directly above the target label (e.g. the predicted label is chocolate and the target label is dark chocolate) or cases where the predicted label does fall under the category directly above the target but not directly (e.g. the predicted label is brown rice (+1 is rice) and the target label is bread (+1 is starchy food) in Table 8.6). When we consider whether the predicted label is equal to or falls under the category directly above the target ('+1 level (lenient)' in Table 8.12), the micro F_1 increases further to 0.498 and the precision increases drastically to 0.861.

When we considered if both target and predicted labels fell under the same overarching category in the hierarchy (i.e. 'Top Category'), we saw another increase in performance to $F_1 = 0.556$. An example would be if the model predicted another food that is not a starchy food such as dairy (See Table 8.6). Although this results in a very general categorization, it may nonetheless be useful to medical researchers, practitioners, and patients interested, for instance, in all edible substances or all lifestyle interventions that patients recommend for a certain ADE.

8.4.4. CASE STUDY ON GIST ADE COPING

For the case study on the entire GIST patient forum, we employed multi-label classification using semantic similarity sentence-BERT as it was the best performing method. Negation detection and relation extraction rely on knowing where in the sentence entities occur, but multi-label classification does not provide this information. Thus, we identified the approximate location of each CS (i.e., each assigned label) as the token in the sentence with the highest similarity to the assigned label.

This resulted in a total of 32,643 strategies of which 3% (1,017) are negated and 4% (1,375) are multi-label strategies. Figure 8.4a shows the ten most prevalent coping strategies mentioned on the forum. Manual analysis indicated that a large portion of these were false positives: They either refer to primary medication (e.g. imatinib); surgery

NER	Entity linking	Micro F1	Micro R	Micro P	Macro F1
None	SemSearch SBERT	0.001 ± 0.001	0.093 ± 0.180	0.001 ± 0.001	0.001 ± 0.001
	NLI SBERT	0.016 ± 0.013	0.018 ± 0.014	0.014 ± 0.012	0.008 ± 0.007
	SemSim SBERT	**0.220** ± 0.011	**0.306** ± 0.010	**0.172** ± 0.014	**0.105** ± 0.010
Oracle NER	+ SemSim SBERT	0.142 ± 0.043	**0.410** ± 0.089	0.086 ± 0.028	0.038 ± 0.015
	+ BioSyn (B)	0.236 ± 0.040	0.258 ± 0.039	**0.217** ± 0.040	**0.084** ± 0.018
	+ BioSyn (S)	**0.241** ± 0.029	0.283 ± 0.030	0.210 ± 0.028	0.083 ± 0.011
NER	+ SemSim SBERT	0.130 ± 0.021	**0.202** ± 0.039	0.097 ± 0.017	0.037 ± 0.008
	+ BioSyn (B)	**0.155** ± 0.017	0.168 ± 0.032	**0.151** ± 0.037	**0.049** ± 0.013
	+ BioSyn (S)	0.144 ± 0.026	0.162 ± 0.009	0.134 ± 0.039	**0.049** ± 0.016

Table 8.11: Results for end-to-end extraction of coping strategies. SBERT: Sentence-BERT, SemSim: Semantic Similarity, SemSearch: Semantic Search, BioSyn (B): BioSyn with BioBERT, BioSyn (S): BioSyn with SapBERT.

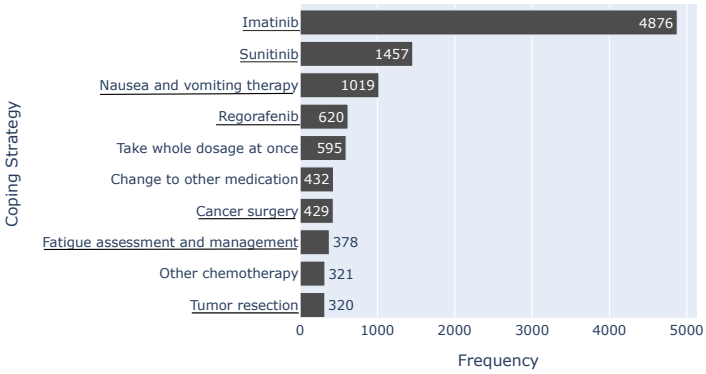
Hierarchy level	Micro F1	Micro R	Micro P	Macro F1
Baseline	0.220 ± 0.011	0.306 ± 0.010	0.172 ± 0.014	0.105 ± 0.010
+1 level (strict)	0.318 ± 0.034	0.336 ± 0.015	0.304 ± 0.048	0.320 ± 0.016
+1 level (lenient)	0.498 ± 0.020	0.350 ± 0.013	0.861 ± 0.063	0.407 ± 0.017
Top categories	0.556 ± 0.018	0.392 ± 0.017	0.952 ± 0.033	0.422 ± 0.040

Table 8.12: Hierarchical evaluation of multi-label semantic similarity SBERT

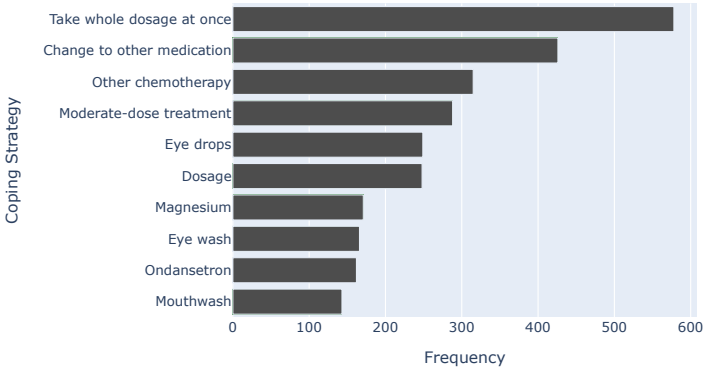
procedures (e.g. cancer surgery) for the disorder itself; side effects (e.g. nausea and vomiting therapy refers to instances of 'nausea'); person names or medical professionals (e.g. oncologist). We manually removed 44 of the 100 most prevalent coping strategies (red lines in Figure 8.4a indicate the removed items in the top-10).

After manual filtering, the total number of coping strategies mentioned was 20.238, of which 3% (694) were negated and 5.5% (1.122) were multi-label. These mentions referred to 2.917 unique coping strategies, which relate to 690 different ADEs. Figure 8.4b shows the most prevalent coping strategies after filtering. Figure 8.5 shows all the coping strategies divided by the highest categories of the ontology (after manual filtering). It appears advice on therapeutic, surgical, or alternative medical procedures ('interventions or procedures' e.g., 'thyroid hormone treatment' or 'moderate-dose treatment') was most prevalent, followed by recommendations to consume medication or supplements and strategies relating to what or how to eat or drink ('eating and drinking').

Figure 8.6 presents the ADEs for which the most coping strategies were provided (See Figure 8.6). The side effect for which the most advice was given was nausea followed by fatigue. In the top 10, various side effects relate to different types of pain (i.e., pain, cramp, painful Mouth) or edema (i.e., edema or periorbital edema). We explored in further detail the most prevalent coping strategies for each of these ADEs. Here we show the results for nausea and cramp, as they most clearly reveal how our semi-automated pipeline can lead to knowledge discovery. We also present results for diarrhea and edema to highlight the problems with negation detection. More analysis for these side effects and the most prevalent coping strategies for the other six side effects are included in Appendix B.



(a) Before manual filtering. Underlined strategies have been manually selected for removal.



(b) After manual filtering.

Figure 8.4: Ten most prevalent coping strategies on the GIST patient forum.

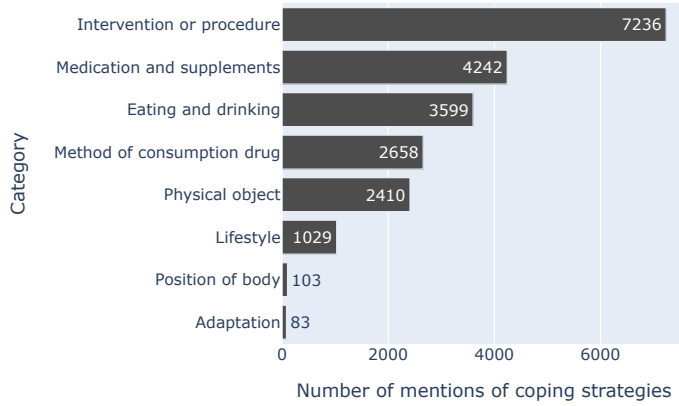


Figure 8.5: Mentions of coping strategies per top category of the ontology (after manual filtering)

Figure 8.7a shows the top 10 coping strategies recommended on the patient forum for nausea. Manual examination of underlying messages reveals that eating and drinking different forms of ginger is recommended, as well as drinking herbal tea (both ginger and peppermint). Patients also recommend taking anti-nausea medication ondansetron and splitting the dosage ('split dosage'). The other categories which relate to how you consume medication (e.g., 'half to one hour before food') do relate to this broader topic, but the specific labels are incorrect. Amongst others, patient recommend to avoid taking medication on an empty stomach and to take it after dinner or just before bed.

Figure 8.7b shows the top 10 coping strategies mentioned on the patient forum for cramps. Manual examination of the underlying messages shows that patients recommend supplements like magnesium, calcium, and potassium ('medication and supplements', 'magnesium', and 'potassium '), food that is high in potassium, tonic water, pickle (juice), and drinking a lot of water ('hydration therapy'). Some patients also recommend exercise ('exercise pain management') although others say it triggers cramps. This is also an example of a case where a coping strategy (exercising) is consistently provided with an incorrect (but semantically similar) label.

Despite decent performance ($F_1 = 0.810$) on our annotated data, qualitative checks revealed that negation detection performed poorly. For instance, manual examination of the underlying messages showed that patients recommend avoiding dairy foods[17] and lactose to reduce diarrhea. However, in Figure 8.8a, only few instances have been negated (red bar) for dairy foods and none for lactose. Another example can be seen in Figure 8.8b, where patients appear divided over whether to avoid or use salt in food ('sodium' and 'low salt food') to reduce edema. The underlying messages, however, are consistent: Patients recommend avoiding salt (blue bar for 'low salt food' and red bar for 'sodium').

[17] The SNOMED concept for dairy is 'dairy foods'

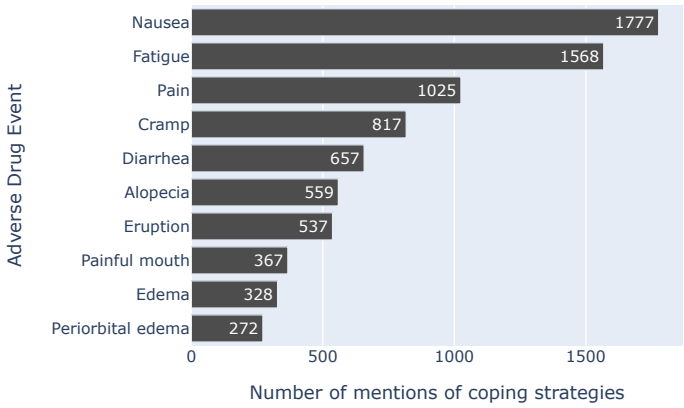


Figure 8.6: The top 10 side effects with the highest number of linked coping strategy mentions (after manual filtering). Alopecia is another term for hair loss, and eruption is another term for rash.

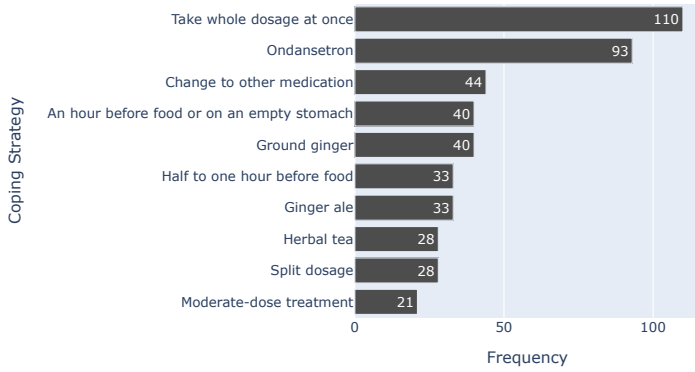
8.5. DISCUSSION

8.5.1. COMPARISON OF APPROACHES

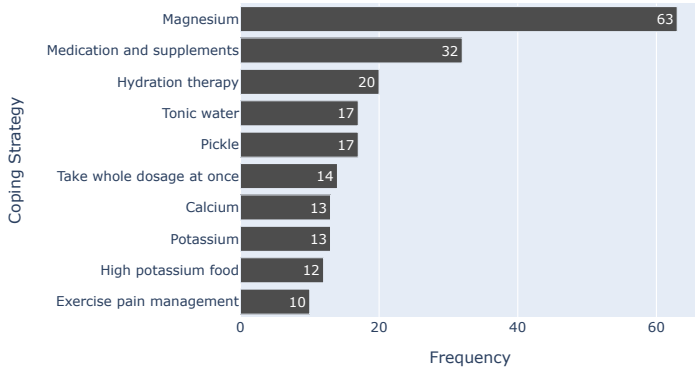
For the extraction of coping strategies for side effects, multi-label classification ($F1$ = 0.220) outperforms named entity recognition (NER) with entity linking (EL) ($F1 = 0.155$). Specifically, Sentence-BERT based on semantic similarity attains the best end-to-end performance, although the quality of the model is still low. Named entity recognition appears to be the bottleneck for the alternative approach, as oracle NER with EL performs even better than multi-label classification ($F1$ = 0.241). This is reflected by the poor token-level NER performance ($F_1 = 0.380$). We found that it is beneficial to include ADE as an additional entity type for NER; This roughly doubled performance (F_1 = 0.200 to $F_1 = 0.380$). Adding a window of one token on each side of the entities further improved performance (to $F_1 = 0.394$), driven by a shift from partially to now fully correct entities. Also, we found that a courser level of ontology matching is considered, the F_1 scores are considerably higher. Overall, we can conclude that multi-label classification is the recommended approach for extracting coping strategies, unless named entity recognition can be improved. One challenge that will remain is the large variety of coping strategy mentions in user-generated text. Increasing the training data will only solve this partly, because there will always be unseen coping strategies in newly seen data.

8.5.2. RELEVANCE OF OUR FINDINGS

These results are also relevant for related tasks, such as the extraction of adverse drug events (ADEs) from social media. Previous work has found that for this task NER is also the bottleneck [159, 193, 194, 335]. Thus, it is worth investigating if multi-label classification is more suited to this task. Moreover, coping strategies for side effects are

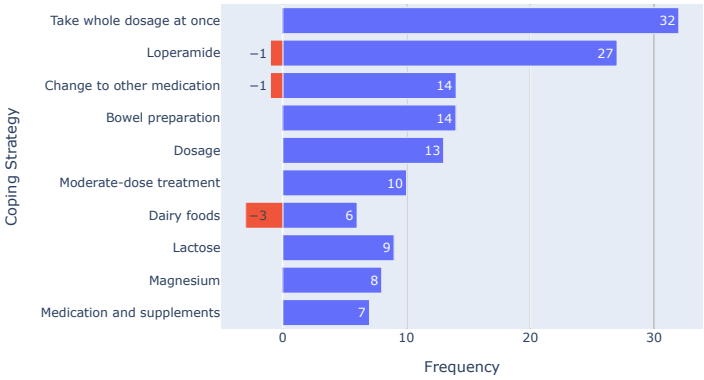


(a) Top 10 coping strategies for nausea

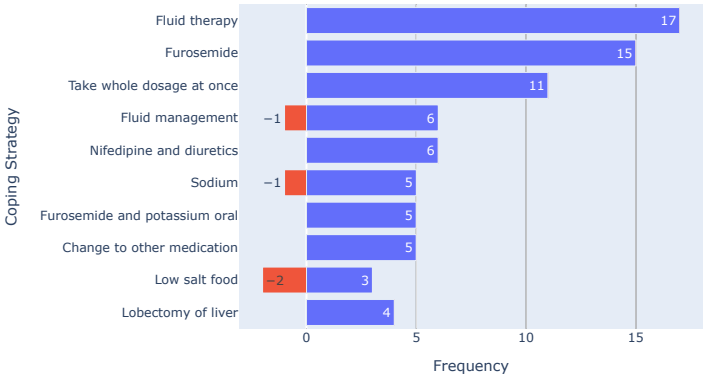


(b) Top 10 coping strategies for cramp

Figure 8.7: Top 10 coping strategies (after manual filtering) without negation



(a) Top 10 coping strategies for diarrhea



(b) Top 10 coping strategies for edema

Figure 8.8: Top 10 coping strategies with negation (after manual filtering). Blue bars indicate that patients recommend taking this strategy and red bars indicate patients that recommend avoiding it (i.e. strategy is negated)

but one type of biomedical complex entity. Unlike named entities, complex entities are often not proper nouns, they tend to be long, and may contain non-entity words (i.e., are discontinuous). Other valuable entities to extract from social media may be advice on psychological coping or coping with the disease in daily life situations e.g. work and childcare. Complex biomedical entities may require different approaches than named entities, and future research is necessary to elucidate whether multi-label classification is consistently preferable to NER with entity linking.

8.5.3. POTENTIAL APPLICATION SETTINGS

Although the quality of our extraction pipeline is insufficient for fully automated knowledge discovery, semi-automated discovery with additional manual qualitative checks can uncover coping strategies for side effects that patients mention online. These can, in turn, be used as input for hypothesis generation. Some examples that we found are drinking ginger tea or taking ondansetron against nausea, and drinking pickle juice or eating potassium-rich food (e.g. bananas) against cramps. Manual examination of the messages underlying a detected strategy can identify cases where the specific label is incorrect (e.g., 'hydration therapy' in Figure 8.7b refers to drinking enough water), as well as cases where it concerns various strategies around a certain topic (e.g., labels referring to how medication should be consumed in Figure 8.7a). These cases likely contribute to the higher performance ($F_1 = 0.498$) when we consider whether the predicted and target labels fall under the same higher order ontological concept.

Expert knowledge is necessary for the manual qualitative checks of the output from the automatic pipeline. Future work could include user studies to estimate the extent of the manual work as well as the extent of the domain knowledge necessary to complete this task. As our work describes the first attempt to tackle this problem, the amount of manual work may also decrease with further improvements to the automatic pipeline. Currently, end-to-end automatic extraction of coping strategies results in a high false positive rate for both MLC and NER+EL. Although recall is more important than precision in a semi-automated system, a high false positive rate is likely to increase the manual work required from experts.

Although we are unable to share our data, we provide the code to visualize and inspect extracted coping strategies[18] in one's own data set. We also share a demonstration of what the visualization would look like.[19] This demonstrates how medical researchers could be aided to conduct adequate qualitative checks and inspect the underlying messages manually using an interface.

Although certain strategies may be self-evident or well known, such as taking anti-nausea medication (e.g., ondansetron) against nausea, others have not been documented previously. Systematic extraction of coping strategies has substantial potential for empowering patients and for generating hypotheses on why these strategies are effective. The coping strategies that are advised should be considered carefully by medical professionals for possible risks before disseminating them amongst patients.

[18] https://github.com/AnneDirkson/CopingStratExtract

[19] https://www.loom.com/share/dda9794a0d354589b95e5b01b5ab23a5

8.5.4. LIMITATIONS

Our work also has a number of limitations. First, the categories included in the ontology are limited to the experiences of GIST patients we collaborated with and the types of coping strategies we encountered on the forum. Although at present our ontology is sufficient to facilitate knowledge discovery, it should be further refined and expanded, for instance through examination of patient forums for other disorders. Furthermore, it would be worthwhile to expand the ontology with categories presented in previous theoretical or qualitative work on coping strategies.

Second, our evaluation of coping strategy extraction is restricted to the labels present in our ground truth data, which cover only 0.6% of the ontology. The performance could thus be overestimated compared to real data if these labels were relatively easy. We preselected discussion threads for annotation based on a high number of patient experiences and at least one drug name using a machine learning model (Chapter 3). Although the performance of this model was good (F_1 = 0.815), discussions around straightforward coping strategies may be easier to identify and thus more likely to be included in the annotated data.

A third limitation is that not all forum posts were subject to coping strategy extraction in the case study. Prior to CS extraction, we selected all posts that contain an ADE and the subsequent 4 posts (see Figure 8.1). Errors in ADE extraction[20] may exclude posts containing coping strategies. Although it may restrict the detected coping strategies, we include this step because previous work has shown that it is beneficial to reduce the data imbalance ratio for extraction [194]. Moreover, our models were trained on similar data. Errors in ADE extraction may also result in the inclusion of posts containing false positives such as symptoms of the disease, resulting in coping strategies that are not directed at resolving adverse drug events.

8.5.5. FUTURE WORK

Aside from further refining our ontology, future work could be directed at exploiting the hierarchical structure of the label space to improve coping strategy extraction, as was done by Rios and Kavuluru [250] and Song et al. [284]. The hierarchical evaluation could also be expanded with more complex hierarchical evaluation metrics such as hierarchical precision and recall [330]. Another possibility would be to include synonyms of the target labels sourced from the UMLS or from the BioPortal term search function. It would also be worthwhile to improve upon our method for ADE–CS relation extraction. Manual error analysis showed that most errors were cases where patients did not explicitly mention which ADE was the target of the coping strategy because it was self-evident to them (e.g. blood pressure medication). Such common sense reasoning appears to often rely on the textual similarity between the ADE and the CS. Thus, relation extraction may be improved by incorporating a similarity metric. Although the performance of negation detection seemed decent (F_1 = 0.810), manual examination of the output revealed negation was not aiding knowledge discovery due to many false positives and negatives. Our heuristics appear insufficient and we recommend future research into improving this module.

Future work could also be directed at researching the low performance of NER for coping strategies, which we expect is due to the descriptive and fuzzy nature of the

[20] ADE extraction has a token-level performance of F_1 0.626 and an entity-level performance of 0.716

entities. We found that the longest correctly identified entity was 9 tokens long, whereas the maximum length of our annotated entities was 29 tokens (see Table 2.3). On average, correctly identified entities were a median of 2 tokens long (± 1 token), partially correctly identified entities were a median of 4 tokens long (± 3 tokens) and missed entities were a median of 2 tokens long (± 3 tokens). It thus appears that missed entities are not on average far longer than correctly identified entities. In contrast, entities that are only partially detected correctly tend to be longer on average. A further investigation of the robustness of NER (e.g. for length and variety of the entities and size of training data) would be insightful for improving the NER model further. Such investigations would also be of interest for other complex entities.

In addition to improving separate modules of the pipeline, future work could include improving their integration. In our current pipeline, the integration of multi-label classification with negation detection and relation extraction was complicated by the need of these modules to know the location of the entity within the sentence. We resolved this by determining the most important token per label that the sentence was labeled with. However, future work could look towards using the attention mechanism of the BERT model underlying multi-label classification, following work on explainable ICD code assignment by Mullenbach et al. [212]. However, this will not be trivial as the Sentence-BERT model is geared towards embedding the entire sentence and does not provide token-specific embeddings. An attention-based approach would also help with differentiating multiple coping strategies (e.g., 'Gatorade, bananas') from a single coping strategy with multiple labels (e.g., 'ginger tea' has the labels 'ginger' and 'herbal tea'). In this work, we defined a coping strategy with two labels as one where the important words were adjacent. Although this is not conventional in related fields such as ICD code detection or ADE extraction, we allow for multiple labels per strategy to curb the exponential growth of the ontology by addition of combined labels.

8.6. CONCLUSION

In this chapter, we have presented a new task, the extraction of coping strategies for side effects from online patient discussions. We developed an ontology for coping strategies, initially tailored to our case of Gastrointestinal Stromal Tumors (GIST), and presented the results for automated extraction method. Moreover, we developed the first pipeline for coping strategy extraction which we use in a case study in which we analyzed an online forum for GIST patients. We showed that automatic extraction of coping strategies for side effects is challenging, with F_1 scores of 0.220 for exact matching to the correct ontology item. We therefore recommend the use of our analysis methods in a semi-automatic fashion in interaction with a human expert to enable the generation of new hypotheses for medical research. Another use would be to discover potentially harmful strategies in the patient-to-patient advice for the purpose of interventions by medical experts.