NTCIR-17
Medical Natural Language Processing for Social media and Clinical texts (MedNLP-SC)

Eiji ARAMAKI, Ph.D. @ NAIST
Yuta NAKAMURA, Ph.D., M.D. @ The University of Tokyo

Organizers

Co-chair (general)

Eiji Aramaki, Ph.D. (NAIST, Japan)

Co-chair (general)

Shoko Wakamiya, Ph.D. (NAIST, Japan)

Co-chair (SM Subtask)

Shuntaro Yada, Ph.D. (NAIST, Japan)

Co-chair (RR Subtask)

Yuta Nakamura, M.D. (The University of Tokyo, Japan)

SM Subtask

Aurélie Névéol, Ph.D. (Université Paris-Saclay, CNRS, LISN, France)

SM Subtask

Patrick Paroubek, Ph.D. (Université Paris-Saclay, CNRS, LISN, France)

SM Subtask

Hui-Syuan Yeh (Université Paris-Saclay, CNRS, LISN, France)

SM Subtask

Pierre Zweigenbaum, Ph.D. (Université Paris-Saclay, CNRS, LISN, France)

SM Subtask

Gabriel Herman Bernardim Andrade (NAIST, Japan)

SM Subtask

Faith Wavinya Mutinda, Ph.D. (NAIST, Japan)

SM Subtask

Tomohiro Nishiyama (NAIST, Japan)

SM Subtask

Lisa Raithel (DFKI, Germany, TU Berlin, Germany, and Université Paris-Saclay, CNRS, LISN, France)

SM Subtask

Akiko Aizawa, Ph.D. (NII, Japan)

RR Subtask

Shouhei Hanaoka, M.D., Ph.D. (The University of Tokyo, Japan)

SM Subtask

Yuji Matsumoto, Ph.D. (RIKEN, Japan)

SM Subtask

Noriki Nishida, Ph.D. (RIKEN, Japan)

SM Subtask

Roland Roller, Ph.D. (DFKI, Germany)

SM Subtask

Philippe Thomas, Ph.D. (DFKI, Germany)

SM Subtask

Cyril Grouin, Ph.D. (Université Paris-Saclay, CNRS, LISN, France)

SM Subtask

Thomas Lavergne, Ph.D. (Université Paris-Saclay, CNRS, LISN, France)

SM Subtask

Hiroki Teranishi, Ph.D. (RIKEN, Japan)

SM Subtask

Narumi Tokunaga (RIKEN, Japan)

SM Subtask

Lis Weiji Kanashiro Pereira Ph.D. (NAIST, Japan)

SM Subtask

Peitao Han (NAIST, Japan)

DFG JST anr®

MedNLP-SC Subtasks

- Social Media Adverse Drug Event detection (SM-ADE)
  - Identify a set of symptoms caused by a drug from short messages written by social media users
  - Social media corpus in Japanese, English, German, and French

Co-chair (general)

Eiji Aramaki, Ph.D. (NAIST, Japan)

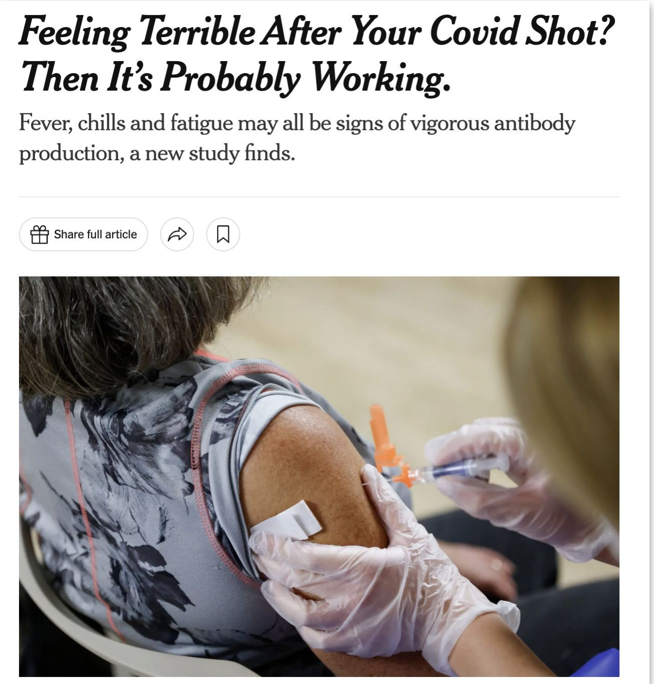
- Radiology Report TNM staging (RR-TNM)
  - Determine the clinical stage of lung cancer from radiology reports, which requires clinical knowledge and complex reasoning
  - Radiology report corpus in Japanese

Co-chair (RR Subtask)

Yuta Nakamura, M.D. (The University of Tokyo, Japan)

SM-ADE Subtask

- Catching Adverse Drug Events (ADE; 副作用) is an important mission with respect to drug safety
- Particularly after COVID-19, people pay much attention to ADEs
- This allows to pick up much information from social media, e.g., X (Twitter) and Facebook
  - We designed a clinical fine grained task
  - BUT....



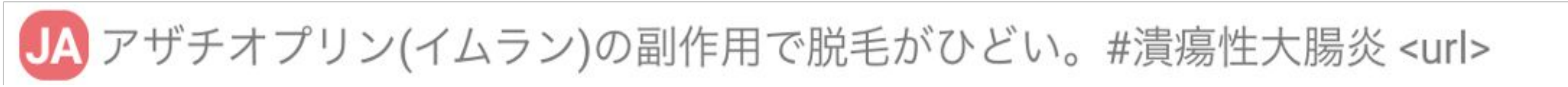
Feeling Terrible After Your Covid Shot? Then It’s Probably Working.

Fever, chills and fatigue may all be signs of vigorous antibody production, a new study finds.

https://www.nytimes.com/2023/10/07/health/covid-vaccine-side-effects.html

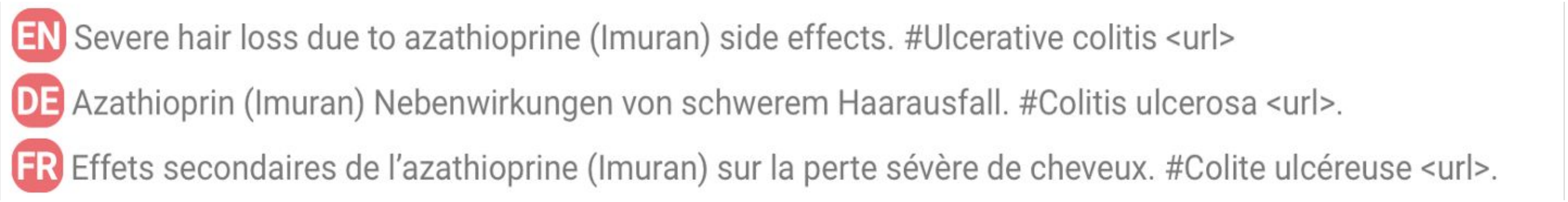
Generated Twitter-like Corpus

Generate 11,000 short messages in Japanese using a pre-trained language model, T5



JA アザチオプリン(イムラン)の副作用で脱毛がひどい。#潰瘍性大腸炎 <url>

Translate the corpus into the other three languages by machine translation with manual check



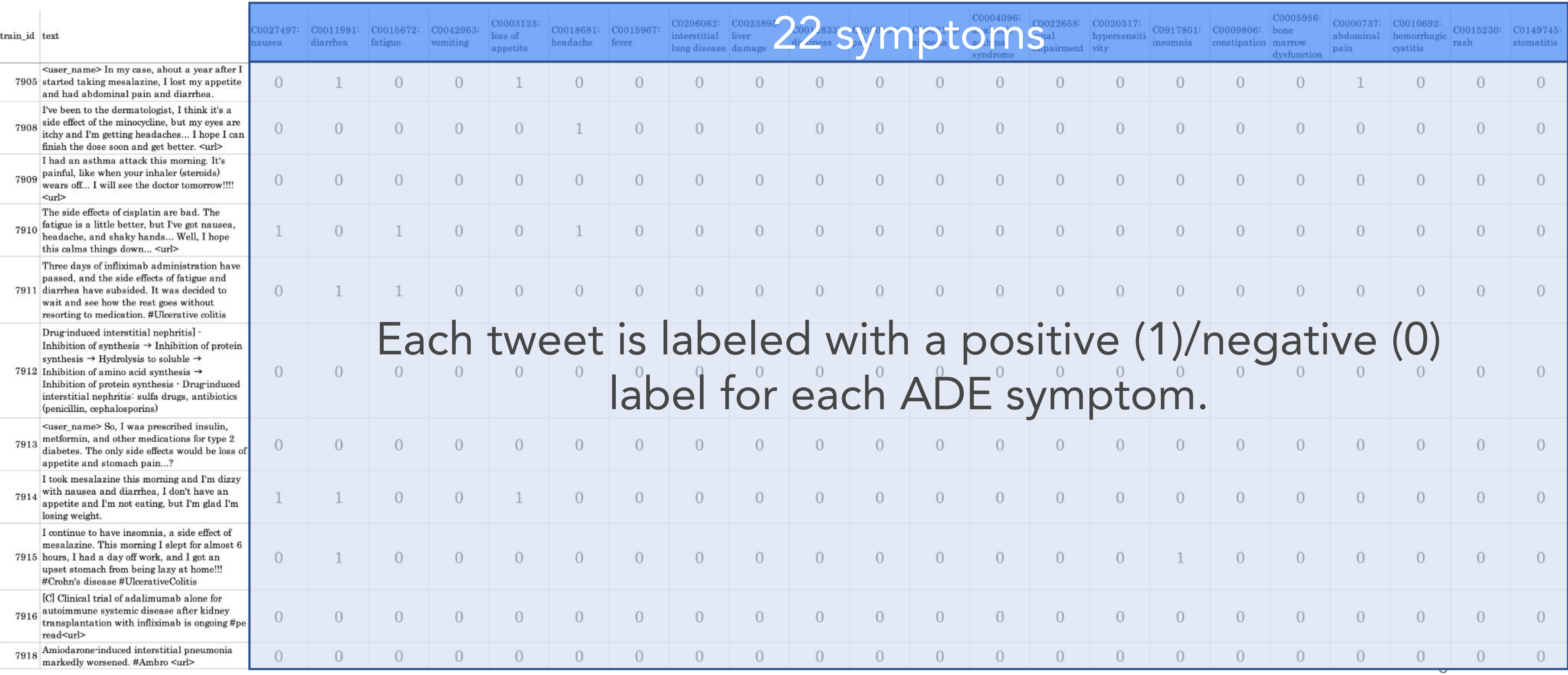
EN Severe hair loss due to azathioprine (Imuran) side effects. #Ulcerative colitis <url>

DE Azathioprin (Imuran) Nebenwirkungen von schwerem Haarausfall. #Colitis ulcerosa <url>.

FR Effets secondaires de l'azathioprine (Imuran) sur la perte sévère de cheveux. #Colite ulcéreuse <url>.

Subtask for Social Media (SM)

Sample data

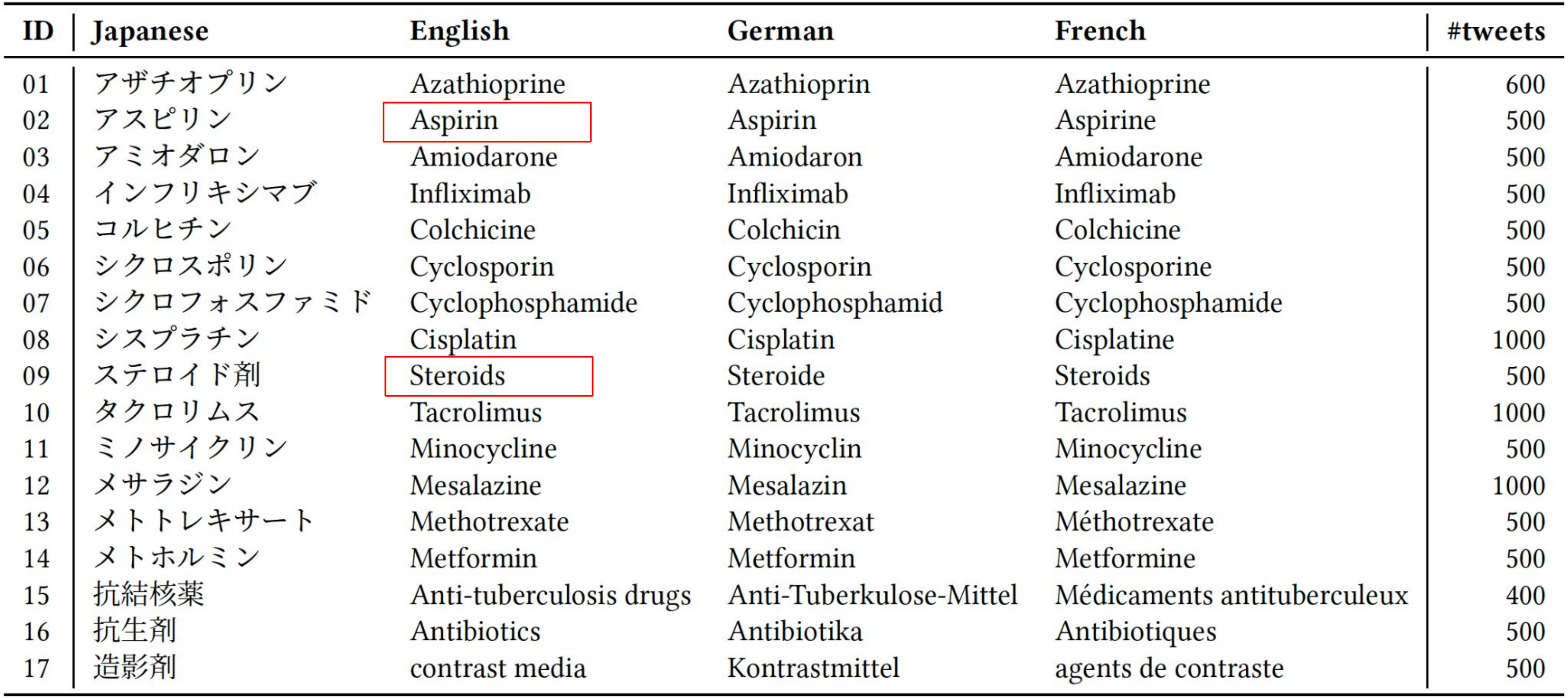


train_id	text	C0027497: nausea	C0011991: diarrhea	C0015672: fatigue	C0042963: vomiting	C0003123: loss of appetite	C0018681: headache	C0015967: fever	C0206062: interstitial lung disease	C0023895: liver damage	C0012833: dizziness	C0030193: pain	C0002170: alopecia	C0004096: analgesic asthma syndrome	C0022658: renal impairment	C0020517: hypersensitivity	C0917801: insomnia	C0009806: constipation	C0005956: bone marrow dysfunction	C0000737: abdominal pain	C0010692: hemorrhagic cystitis	C0015230: rash	C0149745: stomatitis
7905	<user_name> In my case, about a year after I started taking mesalazine, I lost my appetite and had abdominal pain and diarrhea.	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
7908	I've been to the dermatologist, I think it's a side effect of the minocycline, but my eyes are itchy and I'm getting headaches... I hope I can finish the dose soon and get better. <url>	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7909	I had an asthma attack this morning. It's painful, like when your inhaler (steroids) wears off... I will see the doctor tomorrow!!!! <url>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7910	The side effects of cisplatin are bad. The fatigue is a little better, but I've got nausea, headache, and shaky hands... Well, I hope this calms things down... <url>	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7911	Three days of infliximab administration have passed, and the side effects of fatigue and diarrhea have subsided. It was decided to wait and see how the rest goes without resorting to medication. #Ulcerative colitis	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7912	Drug-induced interstitial nephritis] - Inhibition of synthesis → Inhibition of protein synthesis → Hydrolysis to soluble → Inhibition of amino acid synthesis → Inhibition of protein synthesis - Drug-induced interstitial nephritis: sulfa drugs, antibiotics (penicillin, cephalosporins)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7913	<user_name> So, I was prescribed insulin, metformin, and other medications for type 2 diabetes. The only side effects would be loss of appetite and stomach pain...?	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7914	I took mesalazine this morning and I'm dizzy with nausea and diarrhea, I don't have an appetite and I'm not eating, but I'm glad I'm losing weight.	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7915	I continue to have insomnia, a side effect of mesalazine. This morning I slept for almost 6 hours, I had a day off work, and I got an upset stomach from being lazy at home!!! #Crohn's disease #UlcerativeColitis	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
7916	[C] Clinical trial of adalimumab alone for autoimmune systemic disease after kidney transplantation with infliximab is ongoing #pe read<url>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7918	Amiodarone-induced interstitial pneumonia markedly worsened. #Ambro <url>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

22 symptoms

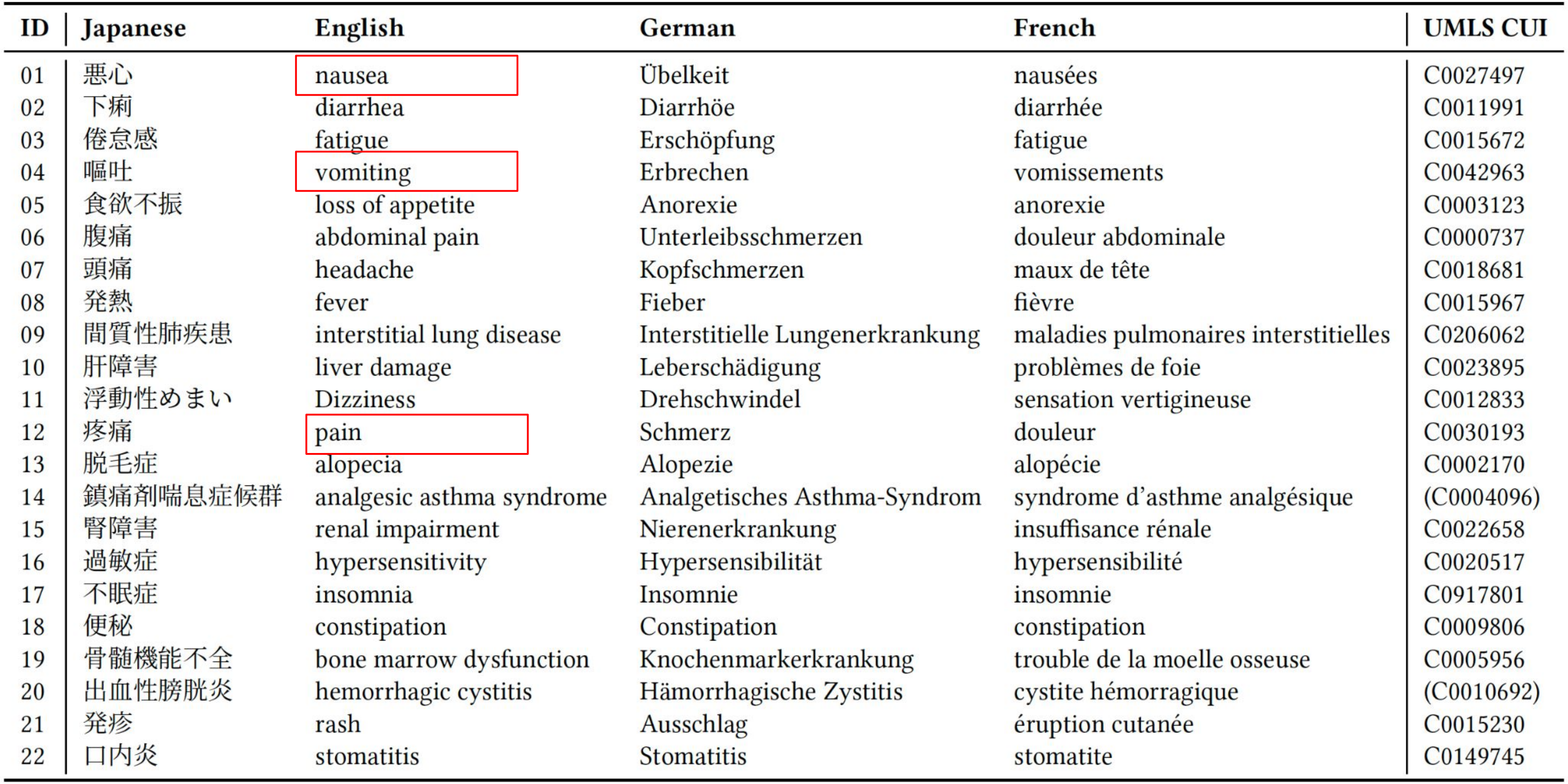
Each tweet is labeled with a positive (1)/negative (0) label for each ADE symptom.

List of 17 Drugs the most popular drugs in social media



ID	Japanese	English	German	French	#tweets
01	アザチオプリン	Azathioprine	Azathioprin	Azathioprine	600
02	アスピリン	Aspirin	Aspirin	Aspirine	500
03	アミオダロン	Amiodarone	Amiodaron	Amiodarone	500
04	インフリキシマブ	Infliximab	Infliximab	Infliximab	500
05	コルヒチン	Colchicine	Colchicin	Colchicine	500
06	シクロスポリン	Cyclosporin	Cyclosporin	Cyclosporine	500
07	シクロフォスファミド	Cyclophosphamide	Cyclophosphamid	Cyclophosphamide	500
08	シスプラチン	Cisplatin	Cisplatin	Cisplatine	1000
09	ステロイド剤	Steroids	Steroide	Steroids	500
10	タクロリムス	Tacrolimus	Tacrolimus	Tacrolimus	1000
11	ミノサイクリン	Minocycline	Minocyclin	Minocycline	500
12	メサラジン	Mesalazine	Mesalazin	Mesalazine	1000
13	メトトレキサート	Methotrexate	Methotrexat	Méthotrexate	500
14	メトホルミン	Metformin	Metformin	Metformine	500
15	抗結核薬	Anti-tuberculosis drugs	Anti-Tuberkulose-Mittel	Médicaments antituberculeux	400
16	抗生剤	Antibiotics	Antibiotika	Antibiotiques	500
17	造影剤	contrast media	Kontrastmittel	agents de contraste	500

List of 22 Symptoms Describing ADEs



ID	Japanese	English	German	French	UMLS CUI
01	悪心	nausea	Übelkeit	nausées	C0027497
02	下痢	diarrhea	Diarrhöe	diarrhée	C0011991
03	倦怠感	fatigue	Erschöpfung	fatigue	C0015672
04	嘔吐	vomiting	Erbrechen	vomissements	C0042963
05	食欲不振	loss of appetite	Anorexie	anorexie	C0003123
06	腹痛	abdominal pain	Unterleibsschmerzen	douleur abdominale	C0000737
07	頭痛	headache	Kopfschmerzen	maux de tête	C0018681
08	発熱	fever	Fieber	fièvre	C0015967
09	間質性肺疾患	interstitial lung disease	Interstitielle Lungenerkrankung	maladies pulmonaires interstitielles	C0206062
10	肝障害	liver damage	Leberschädigung	problèmes de foie	C0023895
11	浮動性めまい	Dizziness	Drehschwindel	sensation vertigineuse	C0012833
12	疼痛	pain	Schmerz	douleur	C0030193
13	脱毛症	alopecia	Alopezie	alopécie	C0002170
14	鎮痛剤喘息症候群	analgesic asthma syndrome	Analgetisches Asthma-Syndrom	syndrome d'asthme analgésique	(C0004096)
15	腎障害	renal impairment	Nierenerkrankung	insuffisance rénale	C0022658
16	過敏症	hypersensitivity	Hypersensibilität	hypersensibilité	C0020517
17	不眠症	insomnia	Insomnie	insomnie	C0917801
18	便秘	constipation	Constipation	constipation	C0009806
19	骨髄機能不全	bone marrow dysfunction	Knochenmarkerkrankung	trouble de la moelle osseuse	C0005956
20	出血性膀胱炎	hemorrhagic cystitis	Hämorrhagische Zystitis	cystite hémorragique	(C0010692)
21	発疹	rash	Ausschlag	éruption cutanée	C0015230
22	口内炎	stomatitis	Stomatitis	stomatite	C0149745

Social Media Corpus

Corpus generation

- Generation of 11,000 short messages in Japanese using a pre-trained language model (T5)
  - Each tweet was manually checked and annotated with a positive (1) or negative (0) label for each ADE symptom
- Translation of the corpus into the other three languages by machine translation (DeepL) with manual check
  - All language share the same symptom label(s)

Four language subsets: JA, EN, DE, and FR

- Each subset consists of 9,957 messages, which were divided into 80% training (7,964 messages) and 20% test (1,993 messages)

Participants

- 8 teams submitted results (+baseline)
- 5 out of 8 teams challenged all languages
- All teams (=8/8) submitted results for EN

- We can see much diversity in participants! :)
  - 8 teams from 7 different countries!
  - all submitting teams are not from Japan

	JA	EN	DE	FR
AILABUD 🇮🇹	✔	✔	✔	✔
FRAG 🇫🇷	✔	✔	✔	✔
HPIDHC 🇩🇪	✔	✔	✔	✔
IMNTPU 🇹🇼		✔		
SRCB 🇨🇳	✔	✔	✔	✔
STIS 🇮🇩		✔		
TMUNLP 🇹🇼		✔		
VLP 🇻🇳	✔	✔	✔	✔

Subtask for Social Media (SM)

Overall Approach

- dedicated binary classification models for each symptom is better than one multilabel classification model (yes: AILABUD, no: HPI)
- trained in all languages with multilingual models (TMUNLP and Baseline)
- additional task for transfer learning, language detection (SRCB)
- data augmentation from other models, e.g., GPT generated samples (HPIDHC), sentiment tags from the sentiment reasoning model VADER (STICS)
- clinical model 📋, e.g. ClinicalBERT, SapBERT
- ensembling models from multiple seeds (most submissions)

	Model	Extra data
AILABUD	SapBERT 📋	no
FRAG	XLM-RoBERTa	no
HPIDHC	GPT-3.5, XLM-RoBERTa	data aug.
IMNTPU	BERT, GPT-3.5, GPT-4	data aug.
SRCB	BERT, XLM-RoBERTa	no
STIS	VADER, BERT	no
TMUNLP	BERT, ClinicalDistilBERT 📋	no
VLP	mBERT, RoBERTa, DeBERTa, XLM-RoBERTa	no
Baseline	BERT, RoBERTa, XLM-RoBERTa	no

Evaluation Metrics

- Full: The performance over ADE labels (0 or 1)
  - Exact Match Accuracy
  - Per ADE Label: Precision, Recall, and $F1$ score for each label (0 and 1) across samples and classes
- Individual: The performance across symptoms
  - Per Symptom Class: Precision, Recall, and $F1$ score for each class
- Binary: How well models can detect examples containing ADEs independent of symptoms

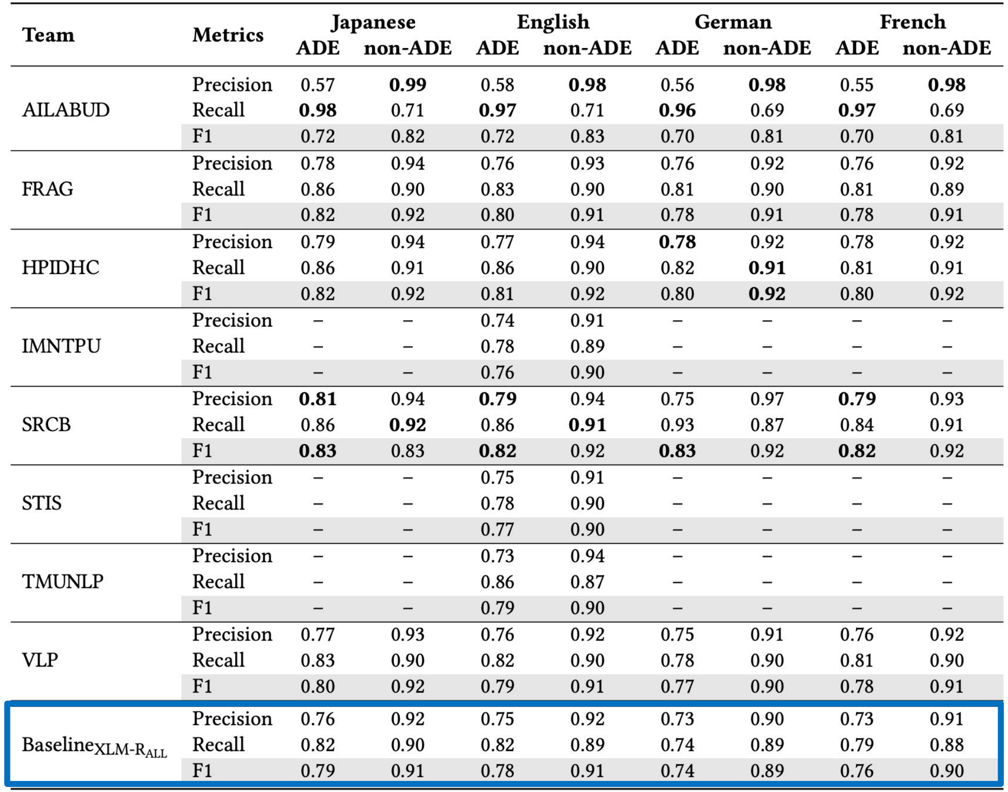
Subtask for Social Media (SM)

Overall on the performance

- Compared to our baseline, F1 improved by around ~5 points; some approaches improved ~10+ points Recall in positive class, which are desirable improvements in medical applications

- Not only the BEST system, but all results are within the range of around 0.8 in F1
  → basic feasibility of the NLP application

Results of the Binary Score for teams in each language track



Team	Metrics	Japanese		English		German		French	
		ADE	non-ADE	ADE	non-ADE	ADE	non-ADE	ADE	non-ADE
AILABUD	Precision	0.57	**0.99**	0.58	**0.98**	0.56	**0.98**	0.55	**0.98**
	Recall	**0.98**	0.71	**0.97**	0.71	**0.96**	0.69	**0.97**	0.69
	F1	0.72	0.82	0.72	0.83	0.70	0.81	0.70	0.81
FRAG	Precision	0.78	0.94	0.76	0.93	0.76	0.92	0.76	0.92
	Recall	0.86	0.90	0.83	0.90	0.81	0.90	0.81	0.89
	F1	0.82	0.92	0.80	0.91	0.78	0.91	0.78	0.91
HPIDHC	Precision	0.79	0.94	0.77	0.94	**0.78**	0.92	0.78	0.92
	Recall	0.86	0.91	0.86	0.90	0.82	**0.91**	0.81	0.91
	F1	0.82	0.92	0.81	0.92	0.80	**0.92**	0.80	0.92
IMNTPU	Precision	–	–	0.74	0.91	–	–	–	–
	Recall	–	–	0.78	0.89	–	–	–	–
	F1	–	–	0.76	0.90	–	–	–	–
SRCB	Precision	**0.81**	0.94	**0.79**	0.94	0.75	0.97	**0.79**	0.93
	Recall	0.86	**0.92**	0.86	**0.91**	0.93	0.87	0.84	0.91
	F1	**0.83**	0.83	**0.82**	0.92	**0.83**	0.92	**0.82**	0.92
STIS	Precision	–	–	0.75	0.91	–	–	–	–
	Recall	–	–	0.78	0.90	–	–	–	–
	F1	–	–	0.77	0.90	–	–	–	–
TMUNLP	Precision	–	–	0.73	0.94	–	–	–	–
	Recall	–	–	0.86	0.87	–	–	–	–
	F1	–	–	0.79	0.90	–	–	–	–
VLP	Precision	0.77	0.93	0.76	0.92	0.75	0.91	0.76	0.92
	Recall	0.83	0.90	0.82	0.90	0.78	0.90	0.81	0.90
	F1	0.80	0.92	0.79	0.91	0.77	0.90	0.78	0.91
$\text{Baseline}_{\text{XLM-R}_{\text{ALL}}}$	Precision	0.76	0.92	0.75	0.92	0.73	0.90	0.73	0.91
	Recall	0.82	0.90	0.82	0.89	0.74	0.89	0.79	0.88
	F1	0.79	0.91	0.78	0.91	0.74	0.89	0.76	0.90

Future Remaining Issues

- How to keep diversity
  - Most submitted systems share the same framework
- After Twitter, social media studies are hard to conduct
  - Some sentences are strange
  - Some sentences are medically dubious

> *Numerous double-blind images were observed in left ventricular block and right ventricular block*
>
> *左心室ブロックおよひ右心室ブロックにおいて多数の* ***二重盲検像****が観察された*

  - Translations often “sound” Japanese, despite being English / German / French

Generated Tweets are not perfect: how to do it better?

How to get the privacy free data?