NTCIR-17 MedNLP-SC Social Media Adverse Drug Event Detection: Subtask Overview

Shoko Wakamiya, Kanashiro Lis, Lisa Raithel, Hui-Syuan Yeh, Peitao Han, Seiji Shimizu, Tomohiro Nishiyama, Gabriel Herman Bernardim Andrade, Noriki Nishida, Hiroki Teranishi, et al.

NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001327

NTCIR-17 MedNLP-SC Social Media Adverse Drug Event Detection: Subtask Overview

Shoko Wakamiya*
NAIST, Japan
wakamiya@is.naist.jp

Lis Kanashiro Pereira*
NAIST, Japan
kanashiro.lis@is.naist.jp

Lisa Raithel*
DFKI Berlin and TU Berlin, Germany
Université Paris-Saclay, CNRS, LISN, France
lisa.raithel@dfki.de

Hui-Syuan Yeh*
Université Paris-Saclay, CNRS, LISN
France
yeh@lisn.fr

Peitao Han*
NAIST, Japan
han.peitao.hr3@is.naist.jp

Seiji Shimizu*
NAIST, Japan
shimizu.seiji.so8@is.naist.jp

Tomohiro Nishiyama
NAIST, Japan
nishiyama.tomohiro.ns5@is.naist.jp

Gabriel Herman Bernardim Andrade
NAIST, Japan
herman_bernardim_andrade.hi1@is.naist.jp

Noriki Nishida
RIKEN, Japan
noriki.nishida@riken.jp

Hiroki Teranishi
RIKEN, Japan
hiroki.teranishi@riken.jp

Narumi Tokunaga
RIKEN, Japan
narumi.tokunaga@riken.jp

Philippe Thomas*
DFKI Berlin, Germany
philippe.thomas@dfki.de

Roland Roller*
DFKI Berlin, Germany
roland.roller@dfki.de

Pierre Zweigenbaum*
Université Paris-Saclay, CNRS, LISN
France
pz@lisn.fr

Yuji Matsumoto
RIKEN, Japan
yuji.matsumoto@riken.jp

Akiko Aizawa
NII, Japan
aizawa@nii.ac.jp

Sebastian Möller
DFKI Berlin and TU Berlin, Germany
sebastian.moeller@tu-berlin.de

Cyril Grouin
Université Paris-Saclay, CNRS, LISN
France
cyril.grouin@lisn.fr

Thomas Lavergne
Université Paris-Saclay, CNRS, LISN
France
lavergne@lisn.fr

Aurélie Névéol
Université Paris-Saclay, CNRS, LISN
France
neveol@lisn.fr

Patrick Paroubek
Université Paris-Saclay, CNRS, LISN
France
pap@lisn.fr

Shuntaro Yada*
NAIST, Japan
s-yada@is.naist.jp

Eiji Aramaki*
NAIST, Japan
aramaki@is.naist.jp

ABSTRACT
This paper presents the Social Media Adverse Drug Event Detection (SM-ADE) subtask as part of the shared task *Medical Natural Language Processing for Social Media and Clinical Texts* (MedNLP-SC) at NTCIR-17. The SM-ADE subtask aims to identify a set of symptoms caused by a drug, referred to as adverse drug event (ADE) detection, within social media texts in multiple languages, including Japanese, English, French, and German. The competition attracted 26 teams, of which eight submitted official runs for the SM-ADE subtask. We believe this task will be essential to develop core technologies of practical medical applications in the near future.

KEYWORDS
Medical Natural Language Processing, Named Entity Recognition, Social Media, Adverse Drug Event

---
* denotes equal contribution

SUBTASKS

SM-ADE-JA
SM-ADE-EN
SM-ADE-DE
SM-ADE-FR

1 INTRODUCTION

Given the rapid progress of natural language processing (NLP) performance, the scope of medical NLP has become broader. Traditionally, medical NLP studies have focused on analyzing texts within hospital settings, such as health records, discharge summaries, and radiology reports [2]. However, in recent times, attention has shifted beyond hospital confines to explore alternative data sources. Among them, social media is one of the most promising resources, since they contain a wealth of direct and personal experiences shared by real patients, making them valuable for medical NLP research [3, 21].

To address this situation, we have proposed a series of Medical Natural Language Processing (MedNLP) tasks so far: MedNLP pilot (NTCIR-10) [24], MedNLP2 (NTCIR-11) [22], MedNLPDoc (NTCIR-12) [1, 23], MedWeb (NTCIR-13) [33, 34], and Real-MedNLP (NTCIR-16) [37]. As a result, our released datasets are widely used in both NLP and biomedical informatics [28, 35]. While the MedNLP series above have mostly focused on Japanese, the last two shared tasks handled multiple languages:

- MedWeb (NTCIR-13) provided pseudo-messages similar to Japanese, English, and Chinese tweets.
- Real-MedNLP (NTCIR-16) offered Japanese-English parallel health records and radiology reports.

Consequently, Real-MedNLP (NTCIR-16) had an increase in the number of overseas participants, with 40% of the teams (four among ten) being from outside Japan.

To further promote this trend, we propose a shared task named Medical Natural Language Processing for Social Media and Clinical Texts (MedNLP-SC) in the context of NTCIR-17. This shared task consists of two subtasks: Social Media Adverse Drug Event Detection (SM-ADE) and Radiology Report TNM staging (RR-TNM) [25]. In this paper, we focus on the SM-ADE subtask. It expands the scope of target languages to include Japanese, English, German, and French.

To develop high-quality multilingual data, we collaborate with researchers from the German Research Center for Artificial Intelligence (DFKI), Germany, and the Interdisciplinary Laboratory of Numerical Sciences at the French National Centre for Scientific Research (LISN, CNRS, Université Paris-Saclay), France.

The SM-ADE subtask aims to identify a set of symptoms caused by a drug from short messages written by social media users. This problem is commonly referred to as adverse drug event detection (henceforth, ADE task). Although we assume that Twitter is the most suitable social media platform, there are ethical and legal concerns about distributing tweets[1]. To deal with this problem, the social media data used in this challenge is artificially generated. The artificial tweets include 17 pre-selected drugs (Table 5) and focus on a set of symptoms. They are generated in Japanese using the pre-trained language model T5 [30]. The resulting tweets are then manually annotated and automatically translated into three other languages: English, French, and German. For instance, given the input sentence: "*Good morning, I have diarrhea like crazy, probably because I'm taking azathioprine, but (...)*", the expected output labels are "diarrhea": positive and "headache": negative. The labels cover 22 symptoms that frequently occur in our corpus, turning the task into a multi-label classification problem. Therefore, this challenge might also be of interest to non-medical NLP groups.

2 TASK SCHEME

The task is represented as a multi-labeling problem for short texts (assuming tweets):

> Input: 1 monolingual text.
> Output: 22 labels. Each label pertains to a symptom and expresses its **positive (1)** or **negative (0)** status as an ADE.

The **positive** label for a symptom indicates a case in which the patient reports a self-experienced ADE with this symptom. The **negative** label for a symptom indicates all other cases, including the case where the symptom is not expressed in the text, or not an ADE, or an ADE the patient does not experience.

In the following artificial examples, only Example 1 receives a **positive** ADE label for *headache*, whereas Examples 2–5 get a **negative** ADE label for *headache*:

(1) I have a *headache* because of Azathioprine.
(2) I have a *headache* which I am treating with Azathioprine.
(3) I don't have an Azathioprine-induced *headache*.
(4) I have a *headache*.
(5) I found an article on Azathioprine-induced *headache*.

In Example 2, the symptom is not an ADE but the reason for the medication. In Example 3, there is no symptom in reality. In Example 4, a patient suffers from a symptom, but it is not described as caused by a drug. Finally, in Example 5, we can find an ADE, but it is just a mention, and the Twitter user did not experience the ADE.

3 MATERIAL

3.1 Overview of the Social Media Corpus

To avoid possible concerns about using social media data such as tweets collected from Twitter (currently X), we generate pseudo-messages similar to Japanese tweets using a pre-trained language model (PLM). The resulting messages are annotated and machine translated into English, French, and German.

We generate 11,000 short messages using T5 [30], one of the standard text-to-text encoder-decoder language models. We adopted its variant pre-trained on Japanese corpora[2].

Each resulting tweet is manually annotated with ADE labels. First, the ADE terms are normalized into MedDRA-Preferred Term (PT) concepts by annotators. Then, to avoid label confusion, we merged several semantically similar MedDRA PTs. For example, *abdominal pain* and *upper abdominal pain* are merged into one category. The lists of the considered ADEs and drugs are provided in Table 4 and Table 5, respectively. The detailed corpus creation process is described in Section 3.2.

[1] https://twitter.com/privacy

[2] https://huggingface.co/sonoisa/t5-base-japanese

Table 1: The overall statistics of the train and test datasets used for the SM-ADE subtask.

Language	Dataset	#samples Total	ADE	non-ADE
Japanese	Train	7,964	2,502	5,462
	Test	1,993	573	1,420
English	Train	7,968	2,506	5,462
	Test	1,993	573	1,420
German	Train	7,974	2,512	5,462
	Test	1,999	579	1,420
French	Train	7,974	2,512	5,462
	Test	1,998	578	1,420

We prepared four language subsets: Japanese (JA), English (EN), German (DE), and French (FR). Each subset consists of 9,957 tweets, divided into 80% training (7,964 tweets) and 20% test set (1,993 tweets). Note that tweets that were difficult to understand by our annotators were removed from every language subset during annotation (38 tweets in total). The labels are the same for each language subset. All drugs are represented in training and test sets except for one drug, which is only present in the test set. This is supposed to simulate the release of a new drug to the public. A summary of the training and test datasets is provided in Tables 1 and 2.

3.2 Corpus Generation

Our synthetic data creation consists of three steps. Each step is detailed below.

Step 1: Generation. First, we collected Japanese tweets (D_{org}) from Twitter. We built the text generation model from the collected tweets to produce Japanese pseudo tweets (D_{gen}).

All of the tweets were collected using 68 drug queries extracted from a Japanese drug-name dictionary[3] and the public Twitter API[4]. During preprocessing, we replaced URLs and user names ("@username") in the original tweets with tags, i.e., `<url>` and `<user_name>` respectively. We further used a Japanese medical named entity recognizer, MedNER-CR-JA[5] [27], to find tweets without any symptom expression and filter them out. Given a sentence, MedNER-CR-JA outputs the sentence with symptoms in the input text tagged with `<C>` and `<CN>`. Note that `<C>` indicates a positive symptom and `<CN>` indicates a negated symptom. Tweets without `<C>` or `<CN>` tags are excluded from the training data.

Based on these filtered original tweets, we fine-tuned T5 to generate synthetic tweets that mention particular drugs. We designed the following prompt for this purpose:

> Input text: "drug_name 使用の Tweet は?<sentinel_0>" (What is a tweet using drug_name?)
> Target text: "<sentinel_0> [original tweet] <sentinel_1>"

[3]https://sociocom.naist.jp/hyakuyaku-dic/
[4]https://developer.twitter.com/en/support/twitter-api
[5]https://huggingface.co/sociocom/MedNER-CR-JA

Table 2: The statistics for the 22 selected symptoms describing ADEs that serve as labels for the multi-label classification. We also show their corresponding Unified Medical Language System [11] Concept Unique Identifier (UMLS CUI) and the number of samples in the train and test datasets containing each symptom. UMLS is a large-scale biomedical knowledge graph containing more than 14M biomedical entity names. Brackets are used when there are no exact matches with English names in the UMLS, and related CUIs are assigned manually. The counts are the same for all language tracks.

ID	UMLS CUI	English Name	#samples Train	Test
01	C0027497	nausea	806	120
02	C0011991	diarrhea	547	136
03	C0015672	fatigue	268	56
04	C0042963	vomiting	193	22
05	C0003123	loss of appetite	249	52
06	C0000737	abdominal pain	354	88
07	C0018681	headache	267	57
08	C0015967	fever	153	53
09	C0206062	interstitial lung disease	16	2
10	C0023895	liver damage	28	2
11	C0012833	dizziness	124	13
12	C0030193	pain	237	72
13	C0002170	alopecia	71	8
14	(C0004096)	analgesic asthma syndrome	95	18
15	C0022658	renal impairment	34	5
16	C0020517	hypersensitivity	184	28
17	C0917801	insomnia	99	34
18	C0009806	constipation	71	31
19	C0005956	bone marrow dysfunction	8	2
20	(C0010692)	hemorrhagic cystitis	11	4
21	C0015230	rash	116	33
22	C0149745	stomatitis	57	22

Here, "<sentinel_0>" and "<sentinel_1>" denote the sentinel tokens used in the pre-training of T5. Using early stopping, fine-tuning was stopped after the 10^{th} epoch.

With the fine-tuned model, for each drug, we generated 10,000 tweets with random sampling and 1,000 tweets with beam search with diversity penalty in order to enhance the diversity within the generated tweets[6].

During post-processing, we filtered out (i) pseudo-messages that do not mention any drug or symptom, (ii) pseudo-messages that are identical to any of the original tweets, and (iii) duplicates. For (i), we applied MedNER-CR-JA to the generated tweets as preprocessing; criterion (ii) is required because the Twitter API policy prohibits the re-distribution of collected tweets.

Step 2: Annotation. Next, we annotated all tweets manually as in the following example:

[6]Since beam search was computationally expensive, we only used it for 1,000 tweets per drug.

aspirin

私は<A>アスピリン喘息</A>があるので、<M>ロキソニン</M>や<M>アセトアミノフェン</M>などはダメなんです (I have <A>aspirin asthma</A>, so no <M>loxonin</M>, <M>acetaminophen</M>, etc.)

The annotation schema is as follows:

<A> positive ADE/ADR
<AN> negative ADE/ADR
<C> positive complaint (non-ADE)
<CN> negative complaint (non-ADE)
<M> drug name

Note that this annotation is not published. It is only used as a preprocessing step for conversion into the final positive and negative labels. For this, we count the number of annotations describing positive ADE mentions (<A>) and take the 22 most frequent ones as labels.

Step 3: Translation. Machine translation (DeepL [7]) was applied to the annotated Japanese pseudo-messages to generate English, German, and French texts. Since the four languages share the same label set, there is no need for manual annotation of all the translated texts. Note, however, that we modified or removed translations that were difficult to understand.

In total, we obtained 9,957 tweets for each language.

3.3 Examples

To clarify the task, this section goes through several examples in the training set.

JA

アザチオプリンを服用して 2ヶ月経ちました。1 週間くらいで全身の発疹はなくなり、かゆみもほぼ無くなっていたのですが、麻疹が少し残ってて怖かったなぁと思います。

EN

I've been on Azathioprine for 2 months now, and after about a week the rash all over my body was gone and the itching was almost gone, but I still had a bit of measles and I think it was scary.

DE

Ich nehme jetzt seit zwei Monaten Azathioprin, und nach etwa einer Woche war der Ausschlag am ganzen Körper verschwunden und der Juckreiz fast weg, aber ich hatte immer noch ein bisschen Masern, und ich glaube, das war beängstigend.

FR

Je prends de l'azathioprine depuis deux mois maintenant, et après environ une semaine, l'éruption cutanée sur tout mon corps avait disparu et les démangeaisons avaient presque disparu, mais j'avais encore un peu de rougeole et je pense que c'était effrayant.

This example shows two symptoms, "measles" and "rash", which correspond to the same label "C0015230 (rash)". Thus, we got one positive label for "rash".

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>

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>.

We can see the disease name "ulcerative colitis" and the symptom "hair loss". As the drug cannot cause ulcerative colitis, we can only attribute a positive label to "alopecia (hair loss)".

JA

<user_name> で、アザチオプリンの血中濃度を調べてきました。やはりステロイド性肝障害が関係してるのかも?血液検査では炎症反応は上がっていたのですが、脱毛症状や発熱には至ってないようです。

EN

<user_name> So I've been checking blood levels of azathioprine. Could it still be related to steroid-induced liver damage? The blood test showed an elevated inflammatory response but did not seem to lead to hair loss symptoms or fever.

DE

<user_name> Also, ich habe die Blutwerte von Azathioprin überprüft. Könnte es immer noch mit einem steroidbedingten Leberschaden zusammenhängen? Der Bluttest zeigte eine erhöhte Entzündungsreaktion, aber es schien nicht zu Haarausfall-Symptomen oder Fieber zu führen.

[7] https://www.deepl.com/translator

FR

<user_name> Donc, j'ai vérifié les niveaux sanguins d'azathioprine. Pourrait-il encore être lié à des lésions hépatiques induites par les stéroïdes ? L'analyse de sang a montré une réponse inflammatoire élevée, mais elle ne semble pas entraîner de symptômes de perte de cheveux ni de fièvre.

In this example, although many symptoms appear, only "liver damage" is mentioned as ADE. Note that "liver damage" is only suspected by the author. We regard such cases as a positive label.

3.4 Data Validation

This section presents our methods to validate the translations and aligned labels between the four languages. First, we designed several methods to check the parallelism of a text and its translation. We use them to flag suspicious translations that should be checked manually.

Length Ratio. This method aims to detect outliers in the translated text pairs.

The ratio of sentence lengths has been used as one of the first methods to check that two sentences are parallel [7]; the length in characters was shown to be more robust than the length in words. However, the characters used to write Japanese have quite different functions than characters used in English, German, or French (Latin script). Moreover, Japanese uses four different writing systems: Kanji (a.k.a. Chinese characters), hiragana and katakana (syllables), and romaji (Latin script), including digits. Counting each of them as one character unit would not be consistent with their widely different information content. As a simple way to mitigate this issue, we transliterated Japanese text using its approximate pronunciation in the Hepburn convention according to the Python pykakasi library[8]. For each text pair (j, f) where j is a Japanese text and f is a foreign text in English, German, or French, we compute the ratio $\frac{l(f)}{l(j)}$ where $l(f)$ is the number of characters in the foreign text and $l(j)$ is the number of characters in the transliterated Japanese text. We checked that this length ratio is approximately normally distributed for each language and computed its mean and standard deviation. We then considered as outliers text pairs whose length ratios were outside $[\min, \max]$ bounds in the corresponding normal distribution, with $(\min, \max) = (0.001, 0.999)$ for English and $(0.010, 0.990)$ for German and French. This resulted in 172 outliers for English, 284 for German, and 279 for French, respectively, summing up outliers below and above the lower and higher bound.

Semantic Similarity. We estimated the semantic similarity between the text pairs using LASER embeddings [31]. We calculated for each text the document embedding and subsequently calculated the cosine similarity between the Japanese source and the respective translations. Data points that fall below the threshold $Q1 - 1.5IQR$ were tagged as outliers. $Q1$ represents the 25th percentile of the data and IQR the interquartile range. This resulted in 292 outliers for English, 313 for German, and 306 for French.

Token Alignment. For each translation pair, we derived the word alignments using SimAlign [13]. We used the proportion of aligned source tokens as a proxy to detect content that was not translated. We flagged examples with suspicious source alignment scores using the same approach as for *Semantic Similarity*. This resulted in 110 outliers for English, 88 for German, and 311 for French.

Back-translation. Each translated document was back-translated into Japanese. We then calculated again the proportion of aligned source tokens with SimAlign and used the same formula as in *Semantic Similarity* to mark outliers. This resulted in 178 outliers for English, 180 for German, and 168 for French.

Manual validation. We aggregated all methods described above for each sample for each language. A manual inspection was conducted for those samples where at least three out of four methods received a flag. This again resulted in 38 outliers for English, 64 for German, and 55 for French. See Figure A.1 in the appendix visualizing the overlapping outliers in the training and test set.

Flagged samples were checked for correctness of translation[9] as well as for correctness of labels. If the label no longer fits the translations, we manually re-translated the example to exclude translation errors. Native Japanese, German, and French speakers verified any changes in labels or translations.

3.5 Remaining Quality Issues

Still, we found several unnatural and medically incorrect expressions in our corpus. We show some examples as follows.

Because I'm using methotrexate, the doctor said "if your arthritis gets worse, you should stop it"
(メトトレキサートを飲んでいるので関節リウマチが酷くなったら中止した方がいいと言われた)

Stopping *methotrexate* due to *arthritis* is not medically correct because *methotrexate* is a drug designed for curing *arthritis*.

Numerous double-blind images were observed in left ventricular block and right ventricular block
左心室ブロックおよび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

The phrase *double-blind images* above sounds like a disease in a human body part. However, since the sub-phrase *double-blind* technically means an experimental procedure in medical research which often involves a placebo, no "images" are thus produced by the procedure at all; in other words, *double-blind* is not at all a radiological method. We regard this type of expression as meaningless.

[8]https://pypi.org/project/pykakasi/, based upon the kakasi library (http://kakasi.namazu.org/index.html.en), which uses the SKK dictionaries (https://skk-dev.github.io/dict/).

[9]Translations were considered "correct" as long as they intuitively made sense, even if grammar or tense was not perfect.

Table 3: Number of systems developed by each team. The teams are sorted alphabetically.

Team	Japanese	English	German	French
AILABUD	2	2	2	2
FRAG	2	2	2	2
HPIDHC	3	3	3	3
IMNTPU	0	3	0	0
SRCB	3	3	3	3
STIS	0	2	0	0
TMUNLP	0	3	0	0
VLP	3	3	3	3
Total	13	21	13	13

4 METHODS

This section briefly introduces the approaches of the eight participating teams that have formally submitted their results, as shown in Table 3. For more information, please refer to the participant system papers for NTCIR-17 MedNLP-SC SM-ADE subtask [6, 8, 10, 16, 18, 26, 29, 32].

AILABUD [29] *JA, EN, DE, and FR*;
This team used multilingual SapBERT [17] across languages via a two-step approach. The first step consisted of binary classification of the pseudo-tweets into the classes ADE versus non-ADE. The second step was an ADE-specific one-versus-rest classification for each symptom. The authors provide models fine-tuned on all languages, but also separately on each language.

FRAG [8] *JA, EN, DE, and FR*;
The authors combine the training data of all four languages and fine-tune the mBERT and the XLM-R [4] base model, with the XLM-R base model performing best. This approach is exactly the same as the XLM-R_ALL baseline system provided by the organizers. The difference in the scores might be due to the choice of hyperparameters, such as random seed and batch size.

HPIDHC [6] *JA, EN, DE, and FR*;
Team HPIDHC employs GPT-3.5-Turbo[10] to generate additional pseudo-tweets in German (approximately 60 for each symptom) and then translate them into Japanese, English, and French to mitigate the label imbalance. They then use XLM-RoBERTa for fine-tuning in French, German, and Japanese and RoBERTa (large) [19] for fine-tuning in English. They further fine-tune one model on all datasets combined. Finally, they compare the models fine-tuned on different dataset combinations, e.g., with/without augmentation, partial augmentation, ensembling of models, and different data splitting and voting methods.

IMNTPU [18] *EN*;
Team IMNTPU applied data augmentation to counteract the imbalance in the classes and compared BioBERT [15], RoBERTa (base and large) [19], GPT 3.5 and GPT 4.0[11] on the English part of the dataset.

SRCB [16] *JA, EN, DE, and FR*;
The authors use BERT and XLM-RoBERTa across all languages and no further external resources.

STIS [32] *EN*;
Team STIS incorporates sentiment features into their processing with the help of a sentiment analysis model, VADER [12]. They use BERT to perform the task on the English data.

TMUNLP [10] *EN*;
The team applies data augmentation [20] and compares BERT (base) and ClinicalDistilBERT fine-tuned on the English part of the data. They further apply Distribution-Balanced Loss [9, 36] during fine-tuning.

VLP [26] *JA, EN, DE, and FR*;
Team VLP compares mBERT, RoBERTa, DeBERTa, and XLM-RoBERTa (large) across languages. They compare two methods of presenting the data to the models: For the first method, they combine all languages in a simple union (vertical) and fine-tune the models. The second method explores a horizontal concatenation of the input data, where one sample corresponds to a combination of four pseudo-tweets (the same one in each language). They then experiment with different feature extraction methods on both data configurations: Sentence vectors produced by a large language model, tf-idf sentence vectors, and a combination of both.

5 EVALUATION

5.1 Evaluation Metrics

Given the large number of labels (22), requiring an exact match of the full set of labels is a very strict evaluation metric. We, therefore, evaluate the labels on two additional levels as follows:

(1) **Full:** We look at the performance over ADE labels (0 or 1).
Exact Match Accuracy Calculates the percentage of exact matches across all samples. The system has to predict the perfect labeling of a sample to be counted; as soon as one symptom is not correctly predicted, the sample will not be counted.
Per ADE Label Calculates precision, recall, and F_1 score for each label (0 and 1) across samples and classes. We provide scores for each label but are mostly interested in those for the positive class since this class is more difficult to predict.

(2) **Individual:** We look at the performance across symptoms.
Per Symptom Class Calculates precision, recall, and F_1 score for each class. This is useful to see if there are any differences in how systems detect different symptoms.

(3) **Binary:** We evaluate how well models can detect examples containing ADEs independent of symptoms. Calculates the performance of classifying a document into the classes "contains ADE" (positive) versus "does not contain ADE" (negative). A document is considered to contain an ADE if at least one symptom class is positive (1). The most interesting scores, in this case, are precision, recall, and F_1 for the

[10]https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates

[11]https://openai.com/gpt-4

positive class. Scores for the positive/negative class are provided.

5.2 Baseline Models

We built several baseline models using our training and test datasets.
Majority Baseline: The majority baseline assigns the zero label (non-ADE) to all test instances since this label is the most commonly occurring category label in the training dataset.
BERT [5]: We fine-tune several BERT base monolingual models, and evaluate each target language. For Japanese, we fine-tune the cl-tohoku/bert-base-japanese-whole-word-masking model. For English, we use the bert-base-uncased model. For German, we use the dbmdz/bert-base-german-uncased model.
RoBERTa [19]: For the French model, we use the camembert-base model, which is based on the RoBERTa base model.
XLM-R [4]: XLM-RoBERTa (XLM-R) is a multilingual version of RoBERTa. It is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. XLM-R has been shown to perform particularly well in low-resource languages, such as Swahili and Urdu. We use the XLM-R base model released by the authors. In this setting, we train and evaluate each language separately (e.g., fine-tune on the English dataset only, and evaluate on the English dataset).
XLM-R$_{\mathbf{ALL}}$: In this setting, we merge the train datasets of all four languages to fine-tune XLM-R and evaluate each language test set.

We used a learning rate of 2×10^{-5} and a batch size 32 in all experiments. The maximum number of epochs was set to 10. We used 0.01 for the weight decay rate and ADAM [14] as our optimizer. We save the best checkpoint during 5-fold validation on the training data. All pre-trained language model implementations used in our experiments are based on the Hugging Face library[12].

5.3 Results

5.3.1 Baseline Results. The baseline results are shown in Tables 6, 7, 8, and 9. Overall, the setting where we merge the train datasets of all four languages (XLM-R$_{\mathrm{ALL}}$) performed best, i.e., fine-tuning XLM-R on multiple languages at the same time leads to better performance.

5.3.2 Results of Participants' Systems. In total, we obtained the outputs of 60 systems submitted by the eight teams: 13 for JA, 21 for EN, 13 for DE, and 13 for FR. Tables 10, 11, 12, 13, 14, 15, and 16 show the detailed results for all participating teams.

The participating teams are classified into two groups: (1) EN group, consisting of the teams that participated only in the English track, and (2) ALL group, consisting of the teams that participated in all four language tracks. While the EN group (IMNTPU, STIS, and TMUNLP) mainly preferred to use monolingual models (e.g., BERT) or a clinical-specific model (e.g., ClinicalDistilBERT), the ALL group preferred to use a cross-language model (e.g., XLM-RoBERTa). However, excluding AILABUD (though its system performance falls within the range of around 0.7), no big difference was observed in the performance of the other teams' systems, as shown in Table 10. AILABUD's system exhibits a trend of high recall but low precision, resulting in a lower F1 score compared to the other teams' systems.

[12]https://huggingface.co/models

Of the eight teams, SRCB achieved the highest performance in all language tracks (0.88 Exact Match Accuracy in JA, 0.87 in EN, 0.86 in DE, and 0.87 in FR), which utilized BERT and XLM-RoBERTa without any additional resources. This indicates that the language-specific technique is not required for this subtask and suggests the feasibility of automatic ADE detection from social media.

6 DISCUSSION

The contributions of this work are described in the following sections.

6.1 Multilingual Medical NLP

Most medical NLP studies have primarily focused on English. Consequently, it is difficult to conduct a comparative analysis with other languages. However, our study, which aimed to create a multilingual medical NLP benchmark in the four languages, demonstrated no big performance change between languages. This indicates that language-specific approaches might not be necessary. However, we do observe a slightly better performance in Japanese (which was the only data that was manually annotated). This difference might be due to possible machine translation errors in the other three languages (English, German, and French). Further analysis is needed on this topic. Finally, we observed that fine-tuning a cross-lingual language model on all languages at the same time overall led to a better performance.

6.2 Difference in Performance across Symptoms

There was considerable variation in the performance observed for each symptom. Symptoms with high frequency in the corpus overall obtained better performance. However, symptoms such as interstitial lung disease, liver damage, bone marrow dysfunction, and hemorrhagic cystitis overall exhibited lower performance, possibly due to the low frequency in the corpus. Further analysis is also needed on this issue.

7 CONCLUSION

This study introduced Social Media Adverse Drug Event Detection (SM-ADE) subtask in the MedNLP-SC, which is a medical NLP shared task handling two different subtasks.

Given the pressing need for NLP solutions not only in our designated tasks but also in numerous medical applications, it is imperative to establish a global framework for organizing and disseminating our approaches and findings. We are confident that our datasets and the approaches and results of all participants will significantly enhance future research endeavors.

Ultimately, the primary contribution of our subtask lies in facilitating discussions and knowledge-sharing among professionals in the field of medical NLP. Given the relatively nascent nature of medical NLP, the community's cohesion remains in its formative stages. Standard corpora and evaluation frameworks are scarce in this domain. Through collaborative efforts led by the task organizers, as well as ongoing discussions between organizers and participants, we anticipate fostering more robust collaborations in the future.

Table 4: The 22 selected symptoms describing ADEs which serve as labels for the multi-label classification. UMLS [11] is a large-scale biomedical knowledge graph containing more than 14M biomedical entity names. Brackets are used when there are no exact matches with English names in the UMLS, and related Concept Unique Identifiers (CUIs) are assigned manually.

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

Table 5: The 17 medication names we used for generating the artificial tweets.

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

CONTRIBUTIONS

Eiji Aramaki, Shoko Wakamiya, and Shuntaro Yada proposed this shared task. Tomohiro Nishiyama, Gabriel Herman Bernardim Andrade, and Seiji Shimizu produced the initial version of the corpus. Peitao Han and Lis Kanashiro Pereira built the baseline systems and evaluated the results. Noriki Nishida, Hiroki Teranishi, Narumi Tokunaga, Yuji Matsumoto, Akiko Aizawa, Sebastian Möller, Thomas Lavergne, and Patrick Paroubek discussed the corpus design. Hui-Syuan Yeh mapped the symptoms to the CUIs. Lisa Raithel, Hui-Syuan Yeh, Roland Roller, Philippe Thomas, Aurélie Névéol, Cyril Grouin, and Pierre Zweigenbaum controlled the quality of the corpus, the label design, the multilingual support and developed the label validation and evaluation scripts.

Table 6: Baseline Results of Exact Match Accuracy.

Baseline	Japanese	English	German	French
Majority		0.71		
BERT	0.80	0.79	0.73	-
RoBERTa	-	-	-	0.71
XLM-R	0.77	0.76	0.72	0.75
XLM-R$_{\mathrm{ALL}}$	**0.84**	**0.83**	**0.80**	**0.81**

Table 7: Baseline Results of the Per ADE Label setting evaluation.

Language	Baseline	Class	Precision	Recall	F1
-	Majority	0	0.98	1.00	0.99
		1	0.00	0.00	0.00
Japanese	BERT	0	0.99	1.00	0.99
		1	0.74	0.69	0.72
	XLM-R	0	0.99	1.00	0.99
		1	0.73	0.52	0.61
	XLM-R$_{\mathrm{ALL}}$	0	**1.00**	1.00	**1.00**
		1	**0.77**	**0.78**	**0.77**
English	BERT	0	0.99	1.00	0.99
		1	**0.74**	0.60	0.66
	XLM-R	0	0.99	1.00	0.99
		1	0.73	0.46	0.57
	XLM-R$_{\mathrm{ALL}}$	0	**1.00**	0.99	0.99
		1	0.73	**0.78**	**0.76**
German	BERT	0	0.98	1.00	0.99
		1	0.67	0.22	0.33
	XLM-R	0	0.99	1.00	0.99
		1	**0.73**	0.24	0.36
	XLM-R$_{\mathrm{ALL}}$	0	0.99	0.99	0.99
		1	0.71	**0.70**	**0.71**
French	RoBERTa	0	0.98	1.00	0.99
		1	**0.72**	0.12	0.20
	XLM-R	0	0.99	1.00	0.99
		1	0.69	0.44	0.54
	XLM-R$_{\mathrm{ALL}}$	0	**1.00**	0.99	0.99
		1	0.71	**0.75**	**0.73**

ACKNOWLEDGEMENTS

The SM-ADE subtask was supported by JST, AIP Trilateral AI Research, Grant Number JPMJCR20G9, JST AIP-PRISM Grant Number JPMJCR18Y1, Japan, as well as ANR grant ANR-20-IADJ-0005-01, France, and DFG grant 442445488, Germany, under the trilateral ANR-DFG-JST AI call.

We thank Dr. Faith Wavinya Mutinda for helping with the translation and the quality check of the translated texts.

Table 8: Baseline Results of the Per Symptom Class setting evaluation (performance across symptoms). Due to space constraints, we omit each symptom class's precision, recall, and F_1 scores.

Language	Baseline	F1 (micro avg.)	F1 (macro avg.)
-	Majority	0.00	0.00
Japanese	BERT	0.72	0.52
	XLM-R	0.61	0.27
	XLM-R$_{\mathrm{ALL}}$	**0.77**	**0.64**
English	BERT	0.66	0.41
	XLM-R	0.57	0.26
	XLM-R$_{\mathrm{ALL}}$	**0.76**	**0.61**
German	BERT	0.33	0.23
	XLM-R	0.36	0.10
	XLM-R$_{\mathrm{ALL}}$	**0.71**	**0.56**
French	RoBERTa	0.20	0.04
	XLM-R	0.54	0.23
	XLM-R$_{\mathrm{ALL}}$	**0.73**	**0.59**

Table 9: Baseline Results of the Binary setting evaluation (ADE vs. non-ADE).

Language	Baseline	Class	Precision	Recall	F1
-	Majority	ADE	0.00	0.00	0.00
		non-ADE	0.71	1.00	0.83
Japanese	BERT	ADE	0.74	0.76	0.75
		non-ADE	0.90	0.89	0.90
	XLM-R	ADE	**0.79**	0.62	0.69
		non-ADE	0.86	**0.93**	0.89
	XLM-R$_{\mathrm{ALL}}$	ADE	0.76	**0.82**	**0.79**
		non-ADE	**0.92**	0.90	**0.91**
English	BERT	ADE	0.75	0.67	0.71
		non-ADE	0.87	0.91	0.89
	XLM-R	ADE	**0.77**	0.54	0.63
		non-ADE	0.83	**0.94**	0.88
	XLM-R$_{\mathrm{ALL}}$	ADE	0.75	**0.82**	**0.78**
		non-ADE	**0.92**	0.89	**0.91**
German	BERT	ADE	0.79	0.35	0.49
		non-ADE	0.79	0.96	0.87
	XLM-R	ADE	**0.80**	0.35	0.49
		non-ADE	0.79	0.96	0.87
	XLM-R$_{\mathrm{ALL}}$	ADE	0.73	**0.74**	**0.74**
		non-ADE	**0.90**	0.89	**0.89**
French	RoBERTa	ADE	**0.86**	0.21	0.33
		non-ADE	0.75	**0.99**	0.85
	XLM-R	ADE	0.76	0.53	0.63
		non-ADE	0.83	0.93	0.88
	XLM-R$_{\mathrm{ALL}}$	ADE	0.73	**0.79**	**0.76**
		non-ADE	**0.91**	0.88	**0.90**

REFERENCES

[1] Eiji Aramaki, Yoshinobu Kano, Tomoko Ohkuma, and Mizuki Morita. 2016. MedNLPDoc: Japanese Shared Task for Clinical NLP. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*. The COLING 2016 Organizing Committee, Osaka, Japan, 13–16. https://aclanthology.org/W16-4203

[2] Eiji Aramaki, Shoko Wakamiya, Shuntaro Yada, and Yuta Nakamura. 2022. Natural Language Processing: from Bedside to Everywhere. *Yearbook of medical informatics* (June 2022).

[3] Yuki Arase, Tomoyuki Kajiwara, and Chenhui Chu. 2020. Annotation of Adverse Drug Reactions in Patients' Weblogs. In *Proceedings of the 12th Language*

Table 10: Results of the Exact Match Accuracy for teams in each language track.

Team	Japanese	English	German	French
AILABUD	0.75	0.71	0.71	0.67
FRAG	0.86	0.84	0.83	0.83
HPIDHC	0.87	0.85	0.85	0.84
IMNTPU	–	0.82	–	–
SRCB	**0.88**	**0.87**	**0.86**	**0.87**
STIS	–	0.82	–	–
TMUNLP	–	0.83	–	–
VLP	0.85	0.84	0.82	0.83
Baseline$_{\text{XLM-R}_{\text{ALL}}}$	0.84	0.83	0.80	0.81

Table 11: Results of the Per ADE Label for teams in each language track.

Team	Metrics	Japanese 0	Japanese 1	English 0	English 1	German 0	German 1	French 0	French 1
AILABUD	Precision	1.00	0.58	1.00	0.51	1.00	0.54	1.00	0.28
	Recall	0.99	**0.97**	0.98	**0.95**	0.98	**0.94**	0.98	**0.94**
	F1	0.99	0.72	0.99	0.66	0.99	00.69	0.99	0.64
FRAG	Precision	1.00	0.77	1.00	0.76	1.00	0.74	1.00	0.73
	Recall	0.99	0.84	0.99	0.83	0.99	0.79	0.99	0.77
	F1	1.00	0.80	1.00	0.79	1.00	0.76	0.99	0.75
HPIDHC	Precision	1.00	**0.79**	1.00	0.76	1.00	**0.77**	1.00	0.76
	Recall	1.00	0.85	0.99	0.83	**1.00**	0.80	1.00	0.76
	F1	1.00	0.82	1.00	0.79	1.00	0.78	1.00	0.77
IMNTPU	Precision	–	–	1.00	0.72	–	–	–	–
	Recall	–	–	0.99	0.76	–	–	–	–
	F1	–	–	0.99	0.74	–	–	–	–
SRCB	Precision	1.00	0.78	1.00	**0.78**	1.00	0.74	1.00	**0.78**
	Recall	1.00	0.87	1.00	0.85	0.99	0.92	1.00	0.84
	F1	1.00	0.82	1.00	**0.81**	1.00	0.82	1.00	**0.84**
STIS	Precision	–	–	0.99	0.76	–	–	–	–
	Recall	–	–	1.00	0.72	–	–	–	–
	F1	–	–	0.99	0.74	–	–	–	–
TMUNLP	Precision	–	–	1.00	0.71	–	–	–	–
	Recall	–	–	0.99	0.83	–	–	–	–
	F1	–	–	0.99	0.76	–	–	–	–
VLP	Precision	1.00	0.76	1.00	0.75	1.00	0.73	1.00	0.73
	Recall	0.99	0.83	0.99	0.80	0.99	0.76	0.99	0.77
	F1	1.00	0.79	1.00	0.78	0.99	0.75	0.99	0.75
Baseline$_{\text{XLM-R}_{\text{ALL}}}$	Precision	1.00	0.77	1.00	0.73	0.99	0.71	1.00	0.71
	Recall	1.00	0.78	0.99	0.78	0.99	0.70	0.99	0.75
	F1	1.00	0.77	0.99	0.76	0.99	0.71	0.99	0.73

Resources and Evaluation Conference. 6769–6776.
[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).
[5] 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).* Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/N19-1423
[6] Smilla Fox, Martin Preiß, Florian Borchert, Aadil Rasheed, and Matthieu-P. Schapranow. 2023. HPIDHC at NTCIR-17 MedNLP-SC: Data Augmentation and Ensemble Learning for Multilingual Adverse Drug Event Detection. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17.* https://doi.org/10.20736/0002001295
[7] William A. Gale and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* 19, 1 (1993), 75–102. https://aclanthology.org/J93-1004
[8] Anubhav Gupta and Frédéric Rayar. 2023. FRAG at the NTCIR-17 MedNLP-SC Task. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17.* https://doi.org/10.20736/0002001279

Table 12: Results of the Binary Score for teams in each language track.

Team	Metrics	Japanese ADE	Japanese non-ADE	English ADE	English non-ADE	German ADE	German non-ADE	French ADE	French 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
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

[9] Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 8153–8161. https://doi.org/10.18653/v1/2021.emnlp-main.643
[10] Yong-Zhen Huang, Yi-Xuan Lin, Eugene Sy, Yu-Lun Hsieh, and Yung-Chun Chang. 2023. TMUNLP at the NTCIR-17 MedNLP-SC Task. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17.* https://doi.org/10.20736/0002001287
[11] Betsy L Humphreys and D_A Lindberg. 1993. The UMLS project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association* 81, 2 (1993), 170.
[12] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media.* 216–225. https://doi.org/10.1609/icwsm.v8i1.14550
[13] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020.* Association for Computational Linguistics, Online, 1627–1643. https://doi.org/10.18653/v1/2020.findings-emnlp.147
[14] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[15] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (Feb. 2020), 1234–1240.
[16] Hongyu Li, Yongwei Zhang, Yuming Zhang, Shanshan Jiang, and Bin Dong. 2023. SRCB at the NTCIR-17 MedNLP-SC Task. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17.* https://doi.org/10.20736/0002001291
[17] Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking. In *Proceedings of ACL-IJCNLP 2021.* 565–574.
[18] Hsiao-Chuan Liu, Vidhya Nataraj, Chia-Tung Tsai, Wen-Hsuan Liao, Tzu-Yu Liu, Mike Tian-Jian Jiang, and Min-Yuh Day. 2023. IMNTPU at the NTCIR-17 Real-MedNLP Task: Multi-Model Approach to Adverse Drug Event Detection from Social Media. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17.* https://doi.org/10.20736/0002001296
[19] 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. *arXiv preprint arXiv:1907.11692* (2019).
[20] Edward Ma. 2019. NLP Augmentation. https://github.com/makcedward/nlpaug.

[21] Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-Garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre, Salvador Lima-López, Ivan Flores, Karen O'Connor, and Others. 2021. Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*. 21–32.

[22] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2014. Overview of the NTCIR-11 MedNLP Task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11, National Center of Sciences, Tokyo*. National Institute of Informatics (NII).

[23] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2016. Overview of the NTCIR-12 MedNLPDoc Task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-12, National Center of Sciences, Tokyo*. National Institute of Informatics (NII).

[24] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, and Eiji Aramaki. 2013. Overview of the NTCIR-10 MedNLP Task. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-10, National Center of Sciences, Tokyo*. National Institute of Informatics (NII).

[25] Yuta Nakamura, Shouhei Hanaoka, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2023. NTCIR-17 MedNLP-SC Radiology Report Subtask Overview: Dataset and Solutions for Automated Lung Cancer Staging. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17*. https://doi.org/10.20736/0002001328

[26] The-Quyen Ngo, Duy-Dao Do, and Phuong Le-Hong. 2023. VLP Methods at the MedNLP-SC Social Media Adverse Drug Event Detection of NTCIR-17. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17*. https://doi.org/10.20736/0002001294

[27] Tomohiro Nishiyama, Mihiro Nishidani, Aki Ando, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2022. NAISTSOC at the NTCIR-16 Real-MedNLP Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*. 330–333.

[28] Sora Ohashi, Junya Takayama, Tomoyuki Kajiwara, Chenhui Chu, and Yuki Arase. 2020. Text Classification with Negative Supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 351–357. https://doi.org/10.18653/v1/2020.acl-main.33

[29] Beatrice Portelli, Alessandro Tremamunno, Simone Scaboro, Emmanuele Chersoni, and Giuseppe Serra. 2023. AILABUD at the NTCIR-17 MedNLP-SC Task: Monolingual vs Multilingual Fine-tuning for ADE Classification. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17*. https://doi.org/10.20736/0002001313

[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[31] Holger Schwenk and Matthijs Douze. 2017. Learning Joint Multilingual Sentence Representations with Neural Machine Translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Vancouver, Canada, 157–167. https://doi.org/10.18653/v1/W17-2619

[32] Lya Hulliyyatus Suadaa, Eko Putra Wahyuddin, and Farid Ridho. 2023. STIS at the NTCIR-17 MedNLP-SC Task: Incorporating Sentiment to Transformer Architecture for Adverse Drug Event Detection on Social Media. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17*. https://doi.org/10.20736/0002001307

[33] Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2017. Overview of the NTCIR-13 MedWeb Task. In *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-13, National Center of Sciences, Tokyo*. National Institute of Informatics (NII).

[34] Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2019. Tweet Classification Toward Twitter-Based Disease Surveillance: New Data, Methods, and Evaluations. *J Med Internet Res* 21, 2 (20 Feb 2019), e12783. https://doi.org/10.2196/12783

[35] Chen-Kai Wang, Onkar Singh, Zhao-Li Tang, and Hong-Jie Dai. 2017. Using a Recurrent Neural Network Model for Classification of Tweets Conveyed Influenza-related Information. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*. Association for Computational Linguistics, Taipei, Taiwan, 33–38. https://aclanthology.org/W17-5805

[36] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-Balanced Loss for Multi-label Classification in Long-Tailed Datasets. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 162–178.

[37] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. 2022. Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-16, National Center of Sciences, Tokyo*. National Institute of Informatics (NII).

Table 13: Results of the Per Symptom Class setting evaluation for all teams in the Japanese track.

Team	Metrics	C0027497	C0011991	C0015672	C0042963	C0003123	C0018681	C0015967	C0206062	C0023895	C0012833	C0030193	C0002170	C0004096	C0022658	C0020517	C0917801	C0009806	C0005956	C0000737	C0010692	C0015230	C0149745
AILABUD	Precision	0.73	0.61	0.67	0.72	0.65	0.74	0.40	**0.33**	0.14	0.48	0.41	0.54	0.45	0.45	0.51	0.62	0.65	0.25	0.62	0.40	0.59	0.65
	Recall	**0.99**	**0.99**	**0.98**	**0.95**	**0.98**	**1.00**	**0.96**	1.00	1.00	0.92	**0.89**	0.88	0.94	**1.00**	**0.93**	**0.91**	0.97	**1.00**	**1.00**	**1.00**	**0.88**	**1.00**
	F1	0.84	0.76	0.80	0.82	0.78	0.85	0.57	0.50	0.25	0.63	0.56	0.67	0.61	0.62	0.66	**0.74**	0.78	0.40	0.76	0.57	0.71	0.79
FRAG	Precision	0.84	0.79	**0.80**	0.82	0.77	0.82	0.72	0.00	0.00	0.65	0.68	0.78	0.94	0.80	0.69	0.72	0.79	0.00	0.77	0.00	0.76	0.60
	Recall	0.94	0.91	0.91	0.82	0.90	0.93	0.74	0.00	0.00	0.85	0.56	0.88	0.89	0.80	0.86	0.68	0.97	0.00	0.89	0.00	0.79	0.82
	F1	0.89	0.85	0.85	0.82	0.83	0.87	**0.73**	0.00	0.00	0.73	0.61	0.82	0.91	0.80	0.76	0.70	0.87	0.00	0.83	0.00	**0.78**	0.69
HPIDHC	Precision	0.82	0.82	0.78	0.87	**0.79**	0.82	**0.73**	0.00	0.50	0.63	0.64	**0.86**	0.94	1.00	0.68	**0.84**	0.79	0.00	**0.80**	**1.00**	**0.77**	0.83
	Recall	0.92	0.94	0.93	0.91	0.92	0.95	0.72	0.00	0.50	0.92	0.53	0.75	0.94	0.80	0.75	0.62	0.97	0.00	0.92	0.50	0.70	0.86
	F1	0.87	0.87	0.85	0.89	0.85	**0.88**	0.72	0.00	0.50	0.75	0.58	0.80	0.94	0.89	0.71	0.71	0.87	0.00	**0.86**	0.67	0.73	0.84
SRCB	Precision	0.83	**0.83**	0.78	0.87	0.78	0.82	0.68	1.00	0.33	**0.67**	0.68	0.78	0.94	1.00	**0.75**	0.81	0.78	0.50	0.76	0.67	0.72	**0.87**
	Recall	0.97	0.91	0.95	0.91	0.94	0.93	0.74	1.00	0.50	0.92	0.69	0.88	0.89	0.80	0.86	0.65	0.94	0.50	0.89	0.50	0.70	0.91
	F1	**0.90**	0.87	0.85	0.89	0.85	0.87	0.71	**1.00**	0.40	**0.77**	**0.69**	0.82	0.91	0.89	**0.80**	0.72	0.85	0.50	0.82	0.57	0.71	**0.89**
VLP	Precision	**0.85**	0.81	0.78	0.82	0.74	0.80	0.69	0.00	**0.67**	0.60	0.58	0.78	0.94	1.00	0.71	0.61	0.76	**1.00**	0.76	0.60	0.76	0.60
	Recall	0.92	0.93	0.95	0.82	0.94	0.93	0.68	0.00	1.00	0.92	0.42	0.88	0.94	0.80	0.79	0.56	0.94	0.50	0.88	0.75	0.67	0.82
	F1	0.88	0.86	0.85	0.82	0.83	0.86	0.69	0.00	**0.80**	0.73	0.48	0.82	0.94	0.89	0.75	0.58	0.84	**0.67**	0.81	0.67	0.71	0.69
$\text{Baseline}_{\text{XLM-R}_{\text{ALL}}}$	Precision	0.84	0.78	0.78	0.85	0.78	0.79	0.67	0.00	1.00	0.64	0.67	0.67	0.89	0.67	0.71	0.65	0.72	0.00	0.79	0.00	0.73	0.68
	Recall	0.91	0.85	0.88	0.77	0.90	0.88	0.57	0.00	0.50	0.69	0.50	0.75	0.94	0.40	0.86	0.50	0.94	0.00	0.86	0.00	0.58	0.77
	F1	0.88	0.81	0.82	0.81	0.84	0.83	0.61	0.00	0.67	0.67	0.57	0.71	0.92	0.50	0.77	0.57	0.82	0.00	0.83	0.00	0.64	0.72

Table 14: Results of the Per Symptom Class setting evaluation for all teams in the English track.

Team	Metrics	C0027497	C0011991	C0015672	C0042963	C0003123	C0018681	C0015967	C0206062	C0023895	C0012833	C0030193	C0002170	C0004096	C0022658	C0020517	C0917801	C0009806	C0005956	C0000737	C0010692	C0015230	C0149745
AILABUD	Precision	0.70	0.52	0.71	0.85	0.65	0.70	0.51	0.33	0.12	0.50	0.36	0.50	0.53	0.36	0.47	0.66	0.68	0.33	0.28	0.50	0.74	0.50
	Recall	**0.99**	**0.95**	**0.98**	**1.00**	**0.92**	**1.00**	**0.96**	**1.00**	**1.00**	**1.00**	**0.83**	0.88	0.94	0.80	**0.93**	**0.85**	0.97	1.00	**0.98**	1.00	**0.85**	**0.95**
	F1	0.82	0.68	0.82	**0.92**	0.76	0.83	0.67	0.50	0.21	0.67	0.51	0.64	0.68	0.50	0.63	**0.74**	0.80	0.50	0.44	0.67	**0.79**	0.66
FRAG	Precision	0.79	0.74	0.76	0.86	0.75	0.82	0.70	0.00	0.00	0.59	0.64	0.70	0.89	0.60	**0.77**	0.73	0.76	0.00	0.75	0.00	0.76	0.65
	Recall	0.92	0.82	0.91	0.82	0.88	0.88	0.66	0.00	0.00	0.77	0.54	0.88	0.94	0.60	0.82	0.71	0.94	0.00	0.86	0.00	0.76	0.77
	F1	0.85	0.78	0.83	0.84	0.81	0.85	0.68	0.00	0.00	0.67	0.59	0.78	0.92	0.60	0.79	0.72	0.84	0.00	0.80	0.00	0.76	0.71
HPIDHC	Precision	0.79	0.78	**0.78**	0.89	0.81	0.82	0.70	0.00	0.00	0.52	**0.68**	0.80	0.85	0.75	0.74	0.79	0.75	0.33	0.73	0.50	0.76	0.68
	Recall	0.94	0.88	0.89	0.77	0.90	0.95	0.74	0.00	0.00	0.85	0.62	**1.00**	0.94	0.60	0.71	0.56	0.87	0.50	0.90	0.50	0.79	0.68
	F1	0.86	0.82	0.83	0.83	0.85	0.88	**0.72**	0.00	0.00	0.65	**0.65**	**0.89**	0.89	0.67	0.73	0.66	0.81	0.40	0.81	0.50	0.78	0.68
IMNTPU	Precision	0.77	0.75	0.76	0.83	0.75	0.77	0.63	0.00	0.00	0.53	0.53	0.70	0.81	0.83	0.76	0.84	0.74	0.33	0.72	0.67	0.75	0.67
	Recall	0.91	0.83	0.89	0.86	0.87	0.95	0.58	0.00	0.00	0.69	0.44	0.88	0.94	**1.00**	0.68	0.47	0.84	0.50	0.75	0.50	0.64	0.64
	F1	0.83	0.79	0.82	0.84	0.80	0.85	0.61	0.00	0.00	0.60	0.48	0.78	0.87	**0.91**	0.72	0.60	0.79	0.40	0.73	0.57	0.69	0.65
SRCB	Precision	0.81	**0.80**	0.76	0.86	**0.82**	0.86	0.65	1.00	**0.50**	**0.69**	0.66	0.78	0.89	**1.00**	0.73	0.79	**0.85**	0.50	0.74	**0.80**	**0.81**	0.82
	Recall	0.96	0.93	0.93	0.86	0.90	0.95	0.74	0.50	0.50	0.85	0.53	0.88	0.89	0.80	0.86	0.56	0.94	0.50	0.91	1.00	0.76	0.82
	F1	0.88	**0.86**	0.84	0.86	**0.86**	**0.90**	0.69	0.67	**0.50**	**0.76**	0.58	0.82	0.89	0.89	0.79	0.66	**0.89**	0.50	**0.82**	**0.89**	0.78	**0.82**
STIS	Precision	0.81	0.78	0.73	0.77	0.81	**0.87**	0.52	0.00	0.00	0.65	0.65	**1.00**	0.89	0.00	0.74	0.80	0.81	0.00	0.73	0.00	0.74	**0.90**
	Recall	0.92	0.76	0.84	0.91	0.85	0.91	0.60	0.00	0.00	0.85	0.46	0.50	0.89	0.00	0.61	0.24	0.68	0.00	0.82	0.00	0.61	0.41
	F1	0.86	0.77	0.78	0.83	0.83	0.89	0.56	0.00	0.00	0.73	0.54	0.67	0.89	0.00	0.67	0.36	0.74	0.00	0.77	0.00	0.67	0.56
TMUNLP	Precision	0.76	0.75	0.74	0.79	0.74	0.79	0.64	1.00	0.00	0.55	0.58	0.70	0.94	0.80	0.67	**0.85**	0.76	0.50	0.62	0.60	0.70	0.64
	Recall	0.91	0.88	0.82	0.86	0.92	0.95	0.68	0.50	0.00	0.85	0.61	0.88	0.94	0.80	0.79	0.50	**1.00**	0.50	0.88	0.75	0.79	0.73
	F1	0.83	0.81	0.78	0.83	0.82	0.86	0.66	0.67	0.00	0.67	0.59	0.78	0.94	0.80	0.72	0.63	0.86	0.50	0.73	0.67	0.74	0.68
VLP	Precision	**0.84**	0.78	0.77	0.81	0.73	0.80	0.70	0.00	0.33	0.52	0.61	0.70	0.94	0.75	0.73	0.72	0.72	**0.67**	0.73	0.40	0.74	0.76
	Recall	0.93	0.88	0.91	0.77	0.90	0.93	0.62	0.00	0.50	0.85	0.51	0.88	0.94	0.60	0.79	0.62	0.84	1.00	0.82	0.50	0.61	0.73
	F1	0.88	0.83	0.84	0.79	0.81	0.86	0.66	0.00	0.40	0.65	0.56	0.78	0.94	0.67	0.76	0.67	0.78	**0.80**	0.77	0.44	0.67	0.74
$\text{Baseline}_{\text{XLM-R}_{\text{ALL}}}$	Precision	0.79	0.73	0.75	**0.90**	0.77	0.79	0.66	0.00	0.00	0.56	0.54	0.70	0.94	0.67	0.71	0.69	0.74	0.00	**0.76**	0.00	0.74	0.64
	Recall	0.93	0.83	0.82	0.86	0.85	0.95	0.55	0.00	0.00	0.77	0.60	0.88	0.94	0.40	0.79	0.53	0.94	0.00	0.82	0.00	0.61	0.73
	F1	0.85	0.78	0.79	0.88	0.81	0.86	0.60	0.00	0.00	0.65	0.57	0.78	0.94	0.50	0.75	0.60	0.83	0.00	0.79	0.00	0.67	0.68

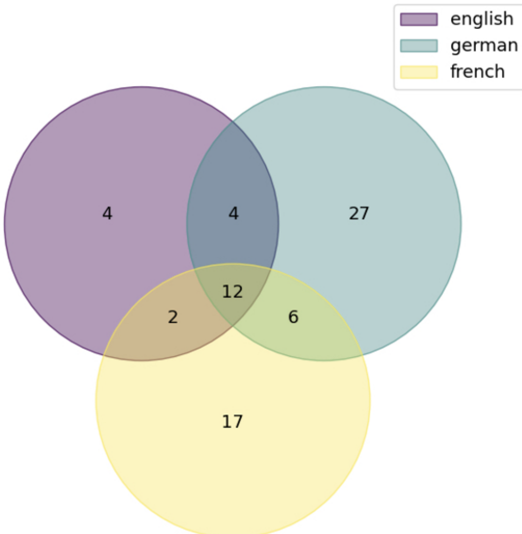
Table 15: Results of the Per Symptom Class setting evaluation for all teams in the German track.

Team	Metrics	C0027497	C0011991	C0015672	C0042963	C0003123	C0018681	C0015967	C0206062	C0023895	C0012833	C0030193	C0002170	C0004096	C0022658	C0020517	C0917801	C0009806	C0005956	C0000737	C0010692	C0015230	C0149745
	Precision	0.65	0.49	0.60	0.62	0.60	0.73	0.46	0.33	0.14	0.59	0.35	0.38	0.55	0.33	0.52	0.60	0.65	0.29	0.60	0.33	0.67	0.51
AILABUD	Recall	0.98	**0.95**	0.96	0.95	0.92	1.00	**0.92**	0.50	1.00	**1.00**	0.82	1.00	0.94	1.00	**0.93**	**0.91**	**0.97**	**1.00**	**0.98**	**1.00**	0.79	0.91
	F1	0.78	0.65	0.74	0.75	0.73	0.84	0.61	0.40	0.25	0.74	0.49	0.55	0.69	0.50	0.67	0.72	0.78	0.44	0.74	0.50	0.72	0.66
	Precision	0.80	0.73	0.76	0.82	0.74	0.83	0.66	0.33	0.00	0.61	0.63	**0.75**	0.89	0.67	**0.77**	0.77	0.77	0.00	0.74	0.00	0.81	0.59
FRAG	Recall	0.92	0.85	0.84	0.82	0.87	0.84	0.58	0.50	0.00	0.85	0.47	0.75	0.89	0.80	0.82	0.71	0.87	0.00	0.85	0.00	0.64	0.77
	F1	0.86	0.79	0.80	0.82	0.80	0.83	0.62	0.40	0.00	0.71	0.54	0.75	0.89	0.73	**0.79**	0.74	0.82	0.00	0.79	0.00	0.71	0.67
	Precision	0.81	0.76	0.76	**0.86**	**0.75**	0.83	**0.73**	1.00	0.00	0.58	**0.66**	0.70	0.89	0.80	0.74	**0.83**	**0.81**	0.50	**0.79**	**0.67**	**0.85**	0.70
HPIDHC	Recall	0.92	0.88	0.89	0.82	0.85	0.91	0.62	0.50	0.00	0.85	0.46	0.88	0.89	0.80	0.71	0.56	0.84	0.50	0.89	0.50	0.67	0.73
	F1	0.86	0.82	0.82	0.84	0.79	0.87	0.67	0.67	0.00	0.69	0.54	0.78	0.89	0.80	0.73	0.67	0.83	0.50	**0.83**	0.57	**0.75**	0.71
	Precision	**0.83**	**0.80**	0.75	0.84	0.73	0.83	0.63	1.00	**0.50**	**0.63**	0.59	0.73	0.74	**1.00**	0.69	0.80	0.74	0.50	0.67	0.60	0.70	**0.83**
SRCB	Recall	0.98	0.94	0.96	0.95	**1.00**	1.00	0.85	**1.00**	1.00	0.92	**0.83**	1.00	0.94	**1.00**	0.89	0.71	0.94	0.50	0.91	0.75	0.79	0.91
	F1	**0.90**	**0.86**	**0.84**	**0.89**	**0.85**	**0.90**	**0.73**	**1.00**	**0.67**	**0.75**	**0.69**	**0.84**	0.83	**1.00**	0.78	**0.75**	0.83	0.50	0.77	**0.67**	0.74	**0.87**
	Precision	0.80	0.72	0.73	0.83	0.73	0.79	0.71	0.00	0.33	0.58	0.55	0.70	0.94	0.80	0.71	0.70	0.73	**1.00**	0.75	0.50	0.75	0.73
VLP	Recall	0.88	0.84	0.84	0.86	0.85	0.88	0.55	0.00	0.50	0.85	0.43	0.88	0.89	0.80	0.71	0.56	0.77	0.50	0.81	0.75	0.55	0.73
	F1	0.84	0.78	0.78	0.84	0.79	0.83	0.62	0.00	0.40	0.69	0.48	0.78	0.91	0.80	0.71	0.62	0.75	**0.67**	0.78	0.60	0.63	0.73
	Precision	0.78	0.75	0.70	0.83	0.70	**0.87**	0.68	0.00	0.00	0.62	0.54	0.64	0.94	0.50	0.69	0.70	0.72	0.00	0.63	0.00	0.69	0.64
Baseline$_{\text{XLM-R}_{\text{ALL}}}$	Recall	0.87	0.70	0.75	0.86	0.73	0.82	0.68	0.00	0.00	0.77	0.44	0.88	0.89	0.20	0.71	0.41	0.84	0.00	0.72	0.00	0.61	0.64
	F1	0.82	0.72	0.72	0.84	0.72	0.85	0.68	0.00	0.00	0.69	0.49	0.74	0.91	0.29	0.70	0.52	0.78	0.00	0.67	0.00	0.65	0.64

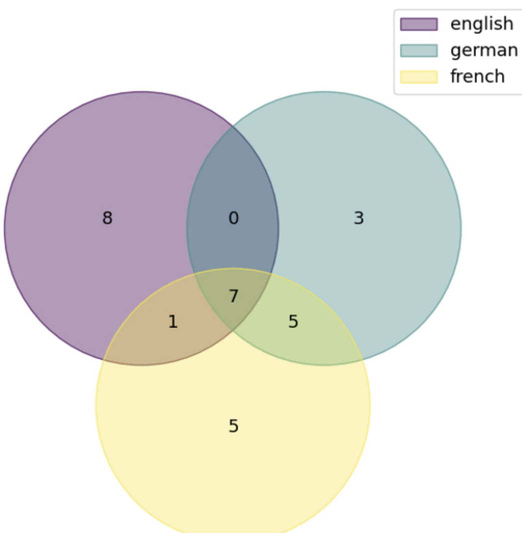
Table 16: Results of the Per Symptom Class setting evaluation for all teams in the French track.

Team	Metrics	C0027497	C0011991	C0015672	C0042963	C0003123	C0018681	C0015967	C0206062	C0023895	C0012833	C0030193	C0002170	C0004096	C0022658	C0020517	C0917801	C0009806	C0005956	C0000737	C0010692	C0015230	C0149745
	Precision	0.69	0.52	0.67	0.73	0.56	0.70	0.47	0.33	**0.13**	0.54	0.37	0.42	0.49	0.33	0.49	0.54	0.67	0.12	0.25	0.44	0.55	0.58
AILABUD	Recall	**1.00**	**0.96**	**0.95**	**1.00**	**1.00**	**0.98**	**0.92**	**1.00**	**1.00**	**1.00**	**0.85**	1.00	0.94	1.00	**0.89**	**0.88**	**0.97**	0.50	**0.94**	1.00	**0.82**	**0.86**
	F1	0.81	0.67	0.79	0.85	0.72	0.82	0.62	0.50	**0.24**	0.70	0.51	0.59	0.64	0.50	0.63	0.67	0.79	0.20	0.39	0.62	0.66	0.69
	Precision	0.82	0.73	0.76	0.83	0.70	0.80	0.67	0.50	0.00	0.55	0.63	0.70	0.89	0.75	0.74	0.71	0.80	0.00	0.68	0.00	0.76	0.64
FRAG	Recall	0.90	0.80	0.86	0.91	0.83	0.82	0.57	0.50	0.00	0.85	0.47	0.88	0.89	0.60	0.82	0.71	0.90	0.00	0.80	0.00	0.79	0.73
	F1	0.86	0.76	0.81	0.87	0.76	0.81	0.61	0.50	0.00	0.67	0.54	0.78	0.89	0.67	**0.78**	**0.71**	0.85	0.00	0.73	0.00	0.78	0.68
	Precision	**0.83**	0.77	0.76	**0.87**	0.73	0.81	**0.74**	0.00	0.00	0.55	0.58	0.78	0.89	1.00	0.75	0.77	**0.84**	0.50	0.70	0.75	**0.86**	0.71
HPIDHC	Recall	0.92	0.88	0.86	0.91	0.83	0.91	0.64	0.00	0.00	0.85	0.43	0.88	0.89	0.60	0.75	0.59	0.84	0.50	0.78	0.75	0.73	0.68
	F1	0.87	0.82	0.81	**0.89**	0.77	0.86	**0.69**	0.00	0.00	0.67	0.50	0.82	0.89	0.75	0.75	0.67	0.84	0.50	0.74	0.75	**0.79**	0.70
	Precision	0.80	**0.81**	0.76	0.83	**0.82**	**0.86**	0.64	1.00	0.00	**0.69**	**0.65**	0.78	0.89	1.00	0.72	**0.79**	0.83	0.50	**0.73**	**0.80**	0.83	**0.82**
SRCB	Recall	0.96	0.91	0.93	0.86	0.90	0.95	0.74	0.50	0.00	0.85	0.49	0.88	0.89	1.00	0.82	0.56	0.94	0.50	0.91	1.00	0.73	0.82
	F1	0.87	**0.86**	0.84	0.84	**0.86**	**0.90**	0.68	0.67	0.00	**0.76**	**0.56**	0.82	0.89	**1.00**	0.77	0.66	**0.88**	0.50	**0.81**	**0.89**	0.77	**0.82**
	Precision	0.80	0.76	**0.78**	0.83	0.72	0.80	0.70	1.00	0.00	0.52	0.59	0.73	0.80	0.75	0.75	0.66	0.75	0.50	0.67	0.60	0.81	0.68
VLP	Recall	0.90	0.82	0.91	0.91	0.85	0.86	0.62	0.50	0.00	0.85	0.46	1.00	0.89	0.60	0.75	0.62	0.77	0.50	0.75	0.75	0.64	0.68
	F1	0.85	0.79	0.84	0.87	0.78	0.83	0.66	0.67	0.00	0.65	0.52	**0.84**	0.84	0.67	0.75	0.64	0.76	0.50	0.71	0.67	0.71	0.68
	Precision	0.78	0.73	0.70	0.79	0.74	0.78	0.70	0.00	0.00	0.53	0.52	0.55	**0.94**	1.00	0.68	0.61	0.70	0.00	0.63	0.00	0.79	0.68
Baseline$_{\text{XLM-R}_{\text{ALL}}}$	Recall	0.91	0.82	0.79	0.68	0.81	0.89	0.66	0.00	0.00	0.69	0.44	0.75	0.94	0.80	0.75	0.56	0.84	0.00	0.76	0.00	0.67	0.68
	F1	0.84	0.78	0.74	0.73	0.77	0.84	0.68	0.00	0.00	0.60	0.48	0.63	**0.94**	0.89	0.71	0.58	0.76	0.00	0.69	0.00	0.72	0.68

APPENDIX



(a) Outliers in the training set.



(b) Outliers in the test set.

Figure A.1: The outliers across languages, showing samples flagged by at least three out of four validation measures.