Social Media Adverse Drug Event Detection (SM-AD) Subtask

Social Media Corpus

- Generated 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
- Translated into the English, German, and French by machine translation (DeepL) with manual check

All language share the same ADE symptom label(s), and each subset consists of 9,957 messages; 80% training (7,946 messages) and 20% test (1,993 messages)

Adverse Drug Event Detection: Subtask Overview

- ADEs independent of language
- Full: The performance over ADE labels (0 or 1)
  - To identify a set of symptoms caused by a drug (ADE; formerly known as "ADRs")
- Individual: The performance across symptoms
- Remaining Issues

Components and Performance

- 5 teams finally submitted their results (+Baseline)
  - 5 of them challenged all languages and all teams challenged the English track
- Most teams assembled models from multiple seeds
- Overall: Desirable improvements in medical applications

Compared to our baseline, F1 improved by around 5%; some improved >10% Recall in positive class

but: All results are within the range of around 0.8 in F1

Results of the Per Symptom Class setting evaluation for all teams in the English track

<table>
<thead>
<tr>
<th>Model</th>
<th>Extra data</th>
<th>Team</th>
<th>Metrics</th>
<th>Japanese</th>
<th>English</th>
<th>German</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>SupBERT</td>
<td>no</td>
<td>AIABRD</td>
<td>Precision 0.57, 0.99</td>
<td>0.58, 0.96, 0.50</td>
<td>0.53</td>
<td>0.86, 0.47</td>
<td>0.86, 0.96</td>
</tr>
<tr>
<td>XLM-RoBERTa</td>
<td>no</td>
<td>FRAG</td>
<td>Precision 0.78, 0.79</td>
<td>0.73, 0.96</td>
<td>0.72, 0.96</td>
<td>0.79, 0.96</td>
<td></td>
</tr>
<tr>
<td>GPT-3, XLM-RoBERTa</td>
<td>data aug</td>
<td>HPDRC</td>
<td>Precision 0.84, 0.84</td>
<td>0.80</td>
<td>0.52, 0.48</td>
<td>0.80, 0.45</td>
<td></td>
</tr>
<tr>
<td>BERT, GPT-3, GPT-4</td>
<td>data aug</td>
<td>MANTPU</td>
<td>Precision 0.80, 0.83</td>
<td>0.82, 0.82</td>
<td>0.82, 0.82</td>
<td>0.80, 0.82</td>
<td></td>
</tr>
<tr>
<td>BERT, XLM-RoBERTa</td>
<td>no</td>
<td>STS</td>
<td>Precision 0.81, 0.93</td>
<td>0.76, 0.91</td>
<td>0.73, 0.90</td>
<td>0.76, 0.93</td>
<td></td>
</tr>
<tr>
<td>VADER, BERT</td>
<td>no</td>
<td>ClinicalDistilBERT</td>
<td>Precision 0.81, 0.94</td>
<td>0.79, 0.94</td>
<td>0.75, 0.97</td>
<td>0.79, 0.93</td>
<td></td>
</tr>
<tr>
<td>mBERT, RoBERTa, DeBERTa, XLM-RoBERTa</td>
<td>no</td>
<td>VLP</td>
<td>Precision 0.80, 0.90</td>
<td>0.81</td>
<td>0.76, 0.80</td>
<td>0.81, 0.76</td>
<td></td>
</tr>
<tr>
<td>BERT, RoBERTa, XLM-RoBERTa</td>
<td>no</td>
<td>Baseline(0.80)</td>
<td>Precision 0.76, 0.82</td>
<td>0.75, 0.93</td>
<td>0.70, 0.95</td>
<td>0.76, 0.79</td>
<td></td>
</tr>
</tbody>
</table>

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
    - Positive (1) or negative (0) label for each ADE symptom

Approaches

If you have any suggestions, please contact us!

Web: https://mednlp.is.naist.jp
mail: mednlp@sc.is.naist.jp
E-mail: mednlp@is.naist.jp

Acknowledgement: This paper was supported by JST, AIP Trilateral AI Research, Grant Number JPMJCR2005, JST AIP Trilateral Grant Number JPMJCR18Y1, Japan, as well as JST grant A012102, PIST, France, and DFG grant 4244608, Germany, under the trilateral AINR-DFGF-IST All.

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

Shoko Wakamia, Li Casashiro Pereira, Lisa Raithel, Katherine Yeh, Pettao Han, Seiji Shimizu, Tomohiro Nishiyama, Gabriel Bernardin Andreade, Noriko Nishida, Hiroki Teranishi, Narumi Tokunaga, Philippe Thomas, Roland Roller, Pierre Zweigenbaum, Yuki Matsumoto, Akiko Aizawa, Sebastian Möller, Cyril Grouin, Thomas Lavergne, Aurélie Névét, Patrick Paroubeck, Shuntaro Yada, Eiji Aramaki
1 NAIST, Japan, 2 DFKI Berlin, Germany, 3 TU Berlin, Germany, 4 Université Paris-Saclay, CNRS, LISN, France, 5 RIKEN, Japan, 6 NII, Japan, 7 JST AI call.