Evaluating the Robustness of Adverse Drug Event Classification Models Using Templates

Dorothea MacPhail[1], **David Harbecke**[1], **Lisa Raithel**[1,2,3], **Sebastian Möller**[1,2]
[1]German Research Center for Artificial Intelligence (DFKI), Berlin
[2]Quality & Usability Lab, Technische Universität Berlin
[3]BIFOLD – Berlin Institute for the Foundations of Learning and Data
[1]{firstname}.{lastname}@dfki.de

Abstract

An adverse drug effect (ADE) is any harmful event resulting from medical drug treatment. Despite their importance, ADEs are often under-reported in official channels. Some research has therefore turned to detecting discussions of ADEs in social media. Impressive results have been achieved in various attempts to detect ADEs. In a high-stakes domain such as medicine, however, an in-depth evaluation of a model's abilities is crucial. We address the issue of thorough performance evaluation in English-language ADE detection with handcrafted templates for four capabilities: Temporal order, negation, sentiment, and beneficial effect. We find that models with similar performance on held-out test sets have varying results on these capabilities.

1 Introduction

When a trained model is applied to real-world data, it may be confronted with phenomena that are under-represented or non-existent in the training data (Belinkov and Bisk, 2019; Moradi and Samwald, 2022). This raises the question of how to evaluate a model's performance and generalization abilities. Reporting summary statistics and held-out test set performance is a common practice in model evaluation. While this can provide an indication of the model's performance and ability to generalize, there are some issues with this practice. Firstly, held-out test sets often arise from the same distribution as the training data and will, therefore, exhibit the same patterns and biases to a high degree. Real-world data, however, may have different feature distribution or exhibit noise. Held-out testing, therefore, often provides an unsatisfactory estimation of a model's performance and generalization abilities (Belinkov and Bisk, 2019; McCoy et al., 2019; Ribeiro et al., 2018).

Secondly, a high model score does not necessarily reveal what the model has learned during training. Research has shown that a model may not learn relevant patterns but instead base its decisions on shallow heuristics or proxies (McCoy et al., 2019). Benchmark challenges have attempted to address this issue by testing models on a wide range of aspects of language (Wang et al., 2019). However, not all aspects can be tested in a benchmark, and the benchmark itself may exhibit unintended biases (Kiela et al., 2021), so the question of what a model has learned remains.

Inspired by the behavioral testing suite CheckList (Ribeiro et al., 2020), we propose the use of template-based test cases to test different capabilities of adverse drug effect (ADE) classification models. ADEs are any harmful consequence to a patient due to medical drug intake. Due to the potential detrimental outcomes of ADEs, the detection of ADEs is an important goal in health-related NLP and has been a subject of research for a considerable time. We test models in understanding of temporal order, positive sentiment, beneficial effects and negation (see Table 1).

In high-stakes domains such as medicine, an in-depth evaluation of a model's abilities is crucial. Related work (Section 2), however, suggests that shortcomings towards selected linguistic phenomena and reliance on proxies for model decisions may exist in models in the biomedical domain.

In this work[1], two transformer-based models for the detection of ADEs in user reports on social media were fine-tuned and tested by conventional held-out testing as well as additional template-based tests. The results of held-out testing and the template-based tests were compared in order to better understand (i) the models' shortcomings and (ii) the potential gaps in knowledge that can occur when a model's abilities are only evaluated via test set performance. We find that models underper-

[1] The templates and code can be found at https://github.com/dfki-nlp/ade_templates

Test Name	Label	Test Description	Example Test Case
Temporal Order standard	no ADE	ADE occurs before drug intake	Before taking cymbalta, I experienced Insomnia.
	ADE	ADE occurs after drug intake	Before having acid reflux, I was put on zoloft.
Temporal Order single time entity	no ADE	ADE occurs before drug intake expressed by a time entity	I was experiencing bad pain in my right arm for 2 weeks, now I started being medicated with Effexor XR.
	ADE	ADE occurs after drug intake expressed by a time entity	3 months ago I started being treated with zoloft, now I started encountering excellerated heart rate.
Temporal Order double time entities	no ADE	ADE occurs before drug intake expressed by two related time entities	3 weeks ago I started suffering from bad pain in my right arm, I have been taking effexor for 2 days.
	ADE	ADE occurs after drug intake expressed by two related time entities	I was enduring Insomnia for 6 weeks, 8 weeks ago I started taking cymbalta.
Positive Sentiment	ADE	ADE occurrence is reported with positive sentiment	I'm taking cymbalta and experiencing cravings for sweets. Still, I am happy my symptoms have reduced.
Beneficial Effect	no ADE	Secondary effect of a drug that is beneficial to the patient	I'm taking Effexor XR and experiencing weight loss. I'm happy because I was trying to lose weight anyway.
	ADE	Secondary effect of a drug that is an ADE as it is not beneficial	For me, weight loss is a side-effect of effexor. It's a problem because I am already underweight.
Negation	no ADE	ADE is negated	I am taking zoloft without suffering from acid reflux.
	ADE	Statement contains negation, ADE is not negated	That's not true, I took zoloft and encountered Insomnia.

Table 1: Overview of all four capabilities tested with example test cases. The temporal order capability has three variations. All test cases have an assigned label, either ADE or no ADE. Filled-in entities are underlined in the example test cases. All test cases are hand-crafted.

form on some capabilities and show differences in some capabilities *despite highly similar F_1-scores on the held-out test set*. We therefore provide the following contributions:

- A curated test bench of 99 templates with 1505 variations to investigate the robustness of ADE classification models across four capabilities.
- A comparison of two popular transformer-based models on long-tail linguistic phenomena in the classification of ADEs.

2 Related Work

Studies on the detection of ADEs in user-generated texts have been conducted since approximately 2010, when Leaman et al. published the first English dataset within this domain. The usual downstream tasks are those common in information extraction: Document classification, to find relevant documents containing mentions of adverse effects; named entity recognition, to identify medication and disease-related mentions; and relation classification, to establish associations between the entity mentions. Approaches for all of these tasks range from rule- and lexicon-based systems (Leaman et al., 2010; Nikfarjam and Gonzalez, 2011) to traditional machine learning pipelines (Gurulingappa et al., 2012; Ginn et al., 2014; Segura-Bedmar et al., 2014) and, recently, deep neural networks (Huynh et al., 2016), specifically transformer-based setups (Weissenbacher et al., 2019; Miftahutdinov et al., 2020; Gusev et al., 2020; Magge et al., 2021b).

However, even advanced models struggle with the supposedly simple task of classifying a document into either "contains an ADE" (henceforth ADE) or "does not contain an ADE" (no ADE), a standard binary classification that is still necessary to find relevant documents for further information extraction. This is often due to a strong class imbalance (in most cases, the documents containing ADEs are in the minority), the usual noise in social media data, ambiguities in health-related statements of patients, and general weaknesses of language models in coping with certain linguistic phenomena not only with respect to ADEs.

For example, Scaboro et al. (2021) have studied the extraction of ADEs from tweets using BERT, SpanBERT (Joshi et al., 2020), and PubMedBERT (Gu et al., 2021). They tested all three models' ability to handle negation and detect shortcomings in all three models. Moradi and Samwald (2022) investigated the robustness of four transformer models specialized in the biomedical and clinical domain over a variety of tasks such as sentence classification, inference, and question answering. The models' robustness is tested by adding minor meaning-preserving changes to the input with the goal of fooling the model. Their findings highlight the vulnerability of state-of-the-art transformer-based models to adversarial input.

Finally, there is CheckList (Ribeiro et al., 2020), a model-agnostic framework aimed at testing a trained model's behavior and gaining an in-depth understanding of its potential shortcomings. CheckList guides the creation of test cases based on natural language *capabilities*, which are used as new inputs to the trained model and subsequently evaluated. The idea is to determine which capabilities (e.g., negation handling, robustness) are necessary for the task the model is intended to perform. Ribeiro et al. (2020) identify three possible test types which can be used for testing the capabilities: the Minimum Functionality Test (MFT), which targets a specific behavior similar to a unit test; the Invariance Test (INV), where the model's robustness to irrelevant perturbations is tested; and the Directional Expectation Test (DIR), which consists of adding perturbations that are expected to lead to a specific outcome. Ribeiro et al. (2020) observe that the CheckList-based evaluation approach could not only uncover bugs in previously tested models but also that CheckList can make the search for bugs more systematic. Recently, updates to CheckList, AdaTest (Ribeiro and Lundberg, 2022) and AdaTest++ (Rastogi et al., 2023), were proposed which assist the user in finding bugs by suggesting topics and test cases in a semi-automated process. While these are valuable additions, we decided to use the template-based approach for this project because we had pre-selected capabilities that we wanted to test with full control over the template design.

CheckList applications include the evaluation of general capabilities of models (Xie et al., 2021) as well as evaluating models in specialized tasks such as offensive speech detection (Bhatt et al., 2021; Manerba and Tonelli, 2021) and automatic text simplification (Cumbicus-Pineda et al., 2021). For the specialized tasks, the authors use CheckList to guide their testing approach by defining new capabilities specific to the task at hand.

In the biomedical and clinical domain, Ahsan et al. (2021) use CheckList to test four linguistic capabilities (negation, temporal order, misspellings, and attributions) on their transformer-based model with a dataset of clinical discharge notes. One of their findings is that the model struggles to correctly distinguish between past and present mentions of substance use in the discharge notes. The detection of ADEs, however, is not part of the research.

The exposure of potential weaknesses in transformer-based models in the biomedical domain motivates an in-depth analysis of models used for ADE detection. To our knowledge, a systematic template-based approach to test model capabilities has not yet been applied to ADE detection.

3 Methods

We use templates to test a selection of linguistic capabilities of binary ADE classification models. To this end, we first manually create templates (see Section 3.1) and then sets of test cases, by using entities to fill placeholders in the templates (see Section 3.3). We then evaluate two fine-tuned classification models on these test cases and compare their predictions with each other and with the models' performance on the held-out test set.

Example 1: Template for Temporal Order (`ADE`)

I started taking `{drug}` before I experienced `{ade}`.

We test four capabilities: *Temporal Order*, *Positive Sentiment*, *Beneficial Effect*, and *Negation* (see Section 3.2). 99 base templates are created with 1505 variations (for details see Table 5 in Appendix A). Each template is also assigned a label (`ADE`/`no ADE`) in accordance with published guidelines for the annotation of ADEs (see Section 4.1.1). The template in Example 1 provides a test case for the capability *Temporal Order* and has a positive label (`ADE`). Filled-in template examples for every capability we test are listed in Table 1. The filled-in templates (test cases) serve as the input to the fine-tuned model for inference. In the following, we present more details about the template creation and the investigated capabilities.

3.1 Template Creation

Template-based evaluation is most effective with a large number of test cases that cover a diverse range of potential inputs. These test cases are based on templates, which include placeholders. For every placeholder, there is a list of potential entity fill-ins as in Example 1, `{drug}` and `{ade}`, which could be filled with, e.g., *Effexor* and *nausea*. The abstraction of test cases to templates allows to systematically capture important linguistic scenarios while creating a large number of different test cases. The process is visualized in Figure 1.

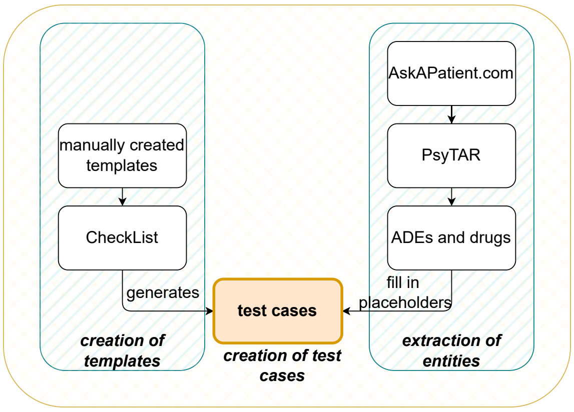

Figure 1: The process for creating test cases.

In the interest of linguistic diversity, variations of base templates were introduced for all capabilities except *Beneficial Effect*. For *Temporal Order* and *Negation* templates, the vocabulary of the base template was modified to increase diversity. *Positive Sentiment* templates underwent syntax variations by exchanging or removing the conjunction between the two phrases.

The templates have a mean token count of 10.6 and 13.4 for the `no ADE` and `ADE` class respectively[2]. After filling in the entities for the placeholders, the average test case length in the experiments is 14.7 for the `no ADE` class and 16.6 for the `ADE` class.

3.2 Capabilities

The choice of capabilities for this work is inspired by considerations on abilities a robust ADE classification model should possess and shortcomings of biomedical models as reported in Section 2. We based the phrasing of the templates on linguistic properties of social media posts: First-person usage, mostly single short sentences, and colloquial language. Contractions were used occasionally. However, usernames, misspellings, and non-standard grammar and punctuation were not applied in the templates as they manifest a separate capability. All templates created can be viewed as templates for a CheckList Minimum Functionality Test (Ribeiro et al., 2020).

To verify the existence of the described phenomena in the dataset, we randomly sampled 1,000 documents and let two annotators check each tweet for the occurrence of these phenomena. The annotations showed that eight of the sampled tweets contain expressions of temporal order, one positive sentiment, one beneficial effect, and one negation. This sample showed that, as expected, the phenomena are rather rare but still exist in the long tail of the data distribution. Nevertheless, an expert would expect a good classification model to have these capabilities.

Temporal Order The templates for testing *Temporal Order* adapt the temporal structure test of Ribeiro et al. (2020) and investigate the model's ability to correctly process information on past, present, and future as expressed in text. In the context of ADE detection, it is important for the model to "understand" temporal order since an effect cannot be an ADE if it occurred before the drug intake. According to the annotation guidelines based on which the data we use for fine-tuning was annotated, an effect occurring after a drug intake was labeled as ADE if the patient draws a connection between the effect and drug intake. Therefore, the templates assume an ADE when a harmful effect occurs after the drug intake.

Positive Sentiment ADEs are often reported using negative sentiment (Alhuzali and Ananiadou, 2019). If many ADE reports contain negative sentiment, an ADE detection model might perform well by using negative sentiment as a proxy. Nevertheless, a report might also be expressed favorably. This could be the case when a patient experiences relief from the original symptoms alongside a mild ADE. Therefore, an ADE detection model should recognize ADEs even when expressed in a positive framing so as not to miss out on less severe ADEs.

Beneficial Effects The third capability is the correct distinction between ADEs and beneficial effects. The latter are secondary effects of a drug that are not related to the reason for using the med-

[2] Tokens were split at whitespace.

ication and which have, nevertheless, a positive outcome for the patient. Note that an effect may be regarded as positive or negative depending on the patient, their general health, and the context. Weight loss, for instance, may be considered a negative secondary drug effect or a beneficial effect depending on the patient. The tests in this work assume that a positive secondary effect is a beneficial effect, not an ADE. The *Beneficial Effect* test that expects a negative class label (`no ADE`) expresses the occurrence of a beneficial effect. The positive class (`ADE`) test consists of test cases that express an ADE that could be classified as a beneficial effect, but the context states that the user views the effect as negative.

Negation *Negation* templates test the model's ability to process negation in text. Negation detection is a general challenge in NLP and a common phenomenon in language (Hossain et al., 2022; Truong et al., 2022). Thus, it is also an important capability for ADE detection. The *Negation* test that expects a negative class label (`no ADE`) contains a negated ADE. The positive class (`ADE`) test cases include an ADE mention as well as a negation without negating the ADE.

3.3 Entity Placeholders

All templates have entity placeholders for a drug name. Templates for *Temporal Order*, *Positive Sentiment*, and *Negation* also have a placeholder for an ADE entity. Templates for *Beneficial Effect* contain an effect that may be considered an ADE or a beneficial effect depending on the context. A list of the effects used in the *Beneficial Effects* tests is provided in Appendix A.2. Template variations of the *Temporal Order* capability that use time entities have placeholders for time expressions. The placeholders are filled with the respective time expressions from a self-created list of entities.

4 Experiments

We frame ADE detection as a binary classification task. We first describe the experiments on the custom dataset and then the experiments on our template-based test cases.

4.1 Fine-Tuning Experiments

The following describes data, training and evaluation on the custom dataset.

Dataset	**#Tweets**	**ADE Ratio (%)**
SMM4H'21 Task 1a	17,426	7.39
SMM4H'17 Task 1	14,880	8.72
NADE	246	0.00
Merged Dataset	**28,468**	**8.75**

Table 2: The number of tweets per dataset and the respective ADE ratio (number of positive samples) of the merged dataset and its three components. 4084 duplicates were removed after merging.

4.1.1 Data

The custom dataset for our experiments consists of three social media corpora: The SMM4H-2021 Shared Task 1a training data (Magge et al., 2021a) (61% of the custom dataset), the SMM4H-2017 Shared Task dataset (Sarker et al., 2018) (38%), and artificially negated tweets from the NADE dataset (Scaboro et al., 2021) (1%), resulting in 28,468 tweets. The data flow and their origin are shown in Figure 2. Dataset statistics are covered in Table 2. In the user-reported texts, each sample either describes an ADE (`ADE`) or does not contain an ADE mention (`no ADE`).

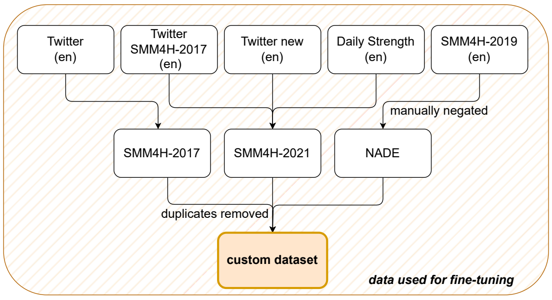

Figure 2: The different data sources for creating the custom dataset for fine-tuning the models.

The SMM4H-2021 Shared Task 1a training data (Magge et al., 2021a) itself consists of posts from Twitter and DailyStrength[3] collected using a list of 81 drugs widespread on the US market (Nikfarjam et al., 2015). The data was annotated by two expert annotators. The annotators did not include beneficial effects in the ADE definition. It further includes some data previously used in the SMM4H-2017 Shared Task (Sarker et al., 2018).

The SMM4H-2017 Shared Task data was collected from Twitter using generic drug names with a total of 250 keywords and subsequently annotated by two annotators. Again, the annotators excluded

[3] www.dailystrength.org

beneficial effects from the ADE definition. Overlapping texts between the SMM4H-2021 data and the SMM4H-2017 data used for our merged custom dataset were removed.

The last part of our custom dataset are artificially negated tweets from the NADE dataset (Scaboro et al., 2021). This dataset consists of tweets originating from the SMM4H-2019 Shared Task (Weissenbacher et al., 2019) and manually negated by annotators. Each negated tweet contains a statement that negates the presence of an ADE. The three components are shown again in Table 2.

We use this merged version of multiple datasets to give the fine-tuning models the best chance to learn different capabilities from varied data. The texts in the custom dataset are between 1 and 34 tokens long.[4] Negative (`no ADE`) samples are slightly shorter on average (16.2 tokens) than positive (`ADE`) samples (18.4 tokens). These are slightly longer than our test cases with an average length of 14.7 tokens and 16.6 tokens. Data splits for training, validation, and testing were created with a 70-10-20 ratio and stratified sampling by class label.

4.1.2 Model Fine-Tuning

For the task of ADE classification, we fine-tune BioRedditBERT (Basaldella et al., 2020) and XLM-RoBERTa (Conneau et al., 2020) on the custom dataset described in Section 4.1.1. BioRedditBERT is a BERT-base uncased model related to BioBERT (Lee et al., 2019), a model pre-trained on the original BERT training corpus (English Wikipedia + BookCorpus) as well as on medical texts sourced from PubMed and PMC. It was then further fine-tuned on a corpus of health-related Reddit posts. XLM-RoBERTa is a popular multilingual model with no specific medical pre-training data. We chose these models to gain insights on robustness of a language model with medical knowledge compared with an general domain language model that has no specific medical knowledge.

The inputs were sampled with replacement weighted by class ratio due to the class imbalance. This sampling strategy resulted in a better F_1-score on the validation dataset.

4.1.3 Held-Out Test Set Evaluation

We evaluate the fine-tuned models on the test set using precision, recall, and F_1-score for each class. The main metric we focused on is F_1 of the positive class due to the large class imbalance. This metric was also used for hyperparameter tuning on the validation set. We compare per-class recall to the models' performances on each capability of the test cases. The goal of this comparison is to determine whether the template-based evaluation approach contradicts the overall impression of the model performance measured by held-out test set performance.

Test	**Label**	**#Test Cases**
Temporal Order standard	no ADE	1,050
	ADE	900
Temporal Order single time entity	no ADE	1,050
	ADE	1,050
Temporal Order double time entities	no ADE	1,575
	ADE	1,575
Positive Sentiment	ADE	2,700
Beneficial Effect	no ADE	120
	ADE	120
Negation	no ADE	825
	ADE	300
Total		**11,265**

Table 3: Number of test cases run per test. We have at least 120 test cases for each capability, so that we can expect our results to be representative.

4.2 Test Case Experiments

We use all templates for each test and randomly select only one template variation per base template for the capabilities *Temporal Order*, *Positive Sentiment*, and *Negation* to have a manageable number of test cases. We created a total of 11,265 test cases, of which 4,620 test cases belong to the negative class (`no ADE`) and 6,645 belong to the positive class (`ADE`). Table 3 shows the number of test cases run per test.

A random sample of 15 ADEs, 15 mild ADEs, 5 drug names, 7 single time entities, and 7 relational time entities was taken. A list of sampled ADEs, mild ADEs, and drug names can be viewed in Appendix B.

4.2.1 Drug and ADE Template Fill-Ins

We need expressions of ADEs and medical drugs to fill in the placeholders in the templates. These are automatically extracted from the PsyTAR dataset (Zolnoori et al., 2019) of patient reports on psychiatric medications. The dataset consists of 891 Ask-a-Patient[5] patient forum posts on the topic of four psychiatric medications: Zoloft, Lexapro,

[4] Tokens were split at white spaces.

[5] www.askapatient.com

Cymbalta, and Effexor XR. The corpus was annotated for ADE mentions by four annotators with a health-related background. A mention was considered an ADE "if there is an explicit report of any sign/symptom that the patient explicitly associated them with the drug consumption" (Zolnoori et al., 2019). All four drug names of PsyTAR were extracted as well as two spelling variations of "Effexor XR" and lowercase versions of all drug names. Statistics on the occurrences of the drug names in the custom training dataset can be found in Table 7 in Appendix B. Extracting ADEs and drug names from the same domain ensures a high likelihood of compatibility between ADEs and medications.

The ADE entities extracted from PsyTAR are user-generated descriptions of ADEs that are often multi-word expressions and which use non-standardized terms. We did not correct grammar and spelling errors in the extracted ADEs.

We created the templates in a way that most short noun phrases[6] fit as ADE entities, therefore, short noun phrases were filtered from the ADE mentions in PsyTAR. A total of 1,227 unique ADEs were extracted, amounting to 36.50% of unique ADE entities in PsyTAR.[7]

For the *Positive Sentiment* test, the extracted ADEs were manually filtered to collect 60 less severe ADEs. This was a necessary step to avoid creating unrealistic test cases such as *"I always have severe pain in my hands when I'm on Cymbalta, but I am happy my symptoms have reduced"*.

The time entities for the variations in *Temporal Order* tests were not extracted but generated. Numbers between 1 and 25 inclusive were combined with a noun (either "days", "weeks", or "months"). A random selection of these combinations was used as time entities.

5 Results

The following presents the results of both the baselines and the template-based test cases.

5.1 Model Baseline

The results of the baseline models can be found in Table 4. All models were evaluated on the same test split of the fine-tuning corpus.

Model	Class	P	R	F$_1$
BioRedditBERT	ADE no ADE	0.720 0.969	0.676 0.975	**0.698** 0.972
XLM-RoBERTa	ADE no ADE	0.720 0.970	0.681 0.975	**0.700** 0.972

Table 4: The results of the baseline models in precision (P), recall (R), and F_1-score on the test split. Positive class F_1 is highlighted as the most popular metric. All scores are very close which would indicate that we can expect similar task understanding of the models.

The F_1-score of BioRedditBERT on the positive class (ADE) is 0.698, whereas XLM-RoBERTa achieves a score of 0.700, which indicates very similar general performance. Due to the large class imbalance, the models reached a higher performance on the majority class (no ADE) with F_1-scores of 0.972. The high overlap in data allows for comparison of this model's performance to the best performing models proposed in the latest SMM4H Shared Task on ADE classification (Weissenbacher et al., 2022).

5.2 Template-Based Test Results

We compare model performance on the custom dataset to each template-based capability test performance separately. Due to the variations in model performance over the two classes, we use per-class recall as a measurement of comparison between the model performance on the custom dataset and the template-based test cases as shown in Figure 3. For both models, all tests with no ADE labels fall short of the baseline performances. The highest level of performance is observed in the *Negation* tests where BioRedditBERT and XLM-RoBERTa pass 92% and 94% of the test cases, respectively. On the other hand, the *Beneficial Effect* tests perform strikingly worse than the baselines with BioRedditBERT and XLM-RoBERTa passing only 7.5% and 5.8% of the test cases, respectively. All three versions of the negative class *Temporal Order* tests lie below the baselines but to a varying degree with a range of 54%-78% for BioRedditBERT and a range of 62%-74% for XLM-RoBERTa.

For ADE, the models perform below the baseline (recall of 68% for both models) on the *standard Temporal Order* and *double time entities Temporal Order* test (25%-48%), while the baseline is exceeded on the *single time entity Temporal Order* test with 90% for BioRedditBERT and 80% for XLM-RoBERTa. Based on the varying model performance on different types of *Temporal Order* tests for both the negative and the positive class,

[6] The longest extracted ADE has a length of 7 tokens.

[7] More details on the extraction can be found in Appendix A.1.

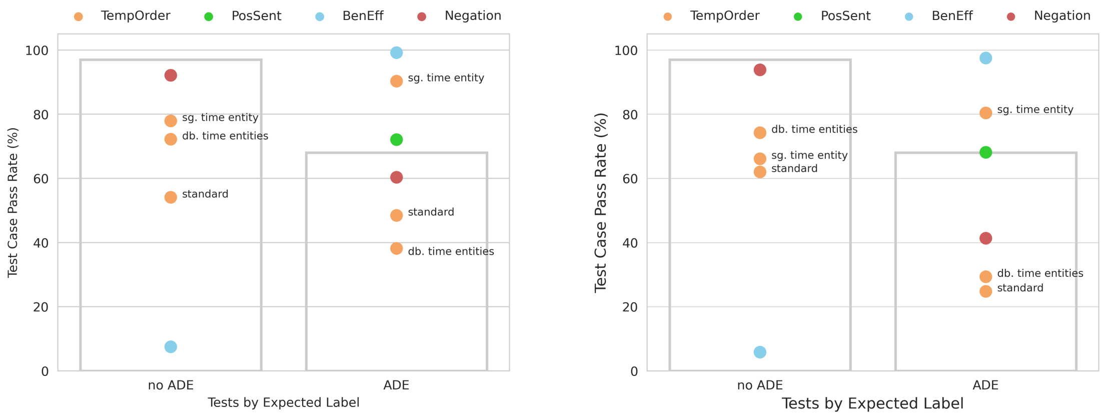

Figure 3: Per-class performance of fine-tuned BioRedditBERT (left) and XLM-RoBERTa (right) on the test set (grey box baseline) and the capability-based test cases. The three distinct types of *Temporal Order* tests refer to variety of *Temporal Order* templates (*standard*, *single* and *double time entity*) highlighted in Table 1. Most test cases are more difficult for the model to solve than the samples from the custom dataset. The biggest difference between the models is the performance on the negation test cases with ADE label, where BioRedditBERT solves 20% more test cases than XLM-RoBERTa. Furthermore, both models have different performances for *Temporal Order* test cases, especially standard cases with ADE label.

one can conclude that the model is not robust to changes in expression of temporal structure: The use of single time entities affects the model performance positively compared to the use of prepositions (*standard Temporal Order*) and double relational time entities. Furthermore, BioRedditBERT (48%) performs much better on *standard Temporal Order* tests than XLM-RoBERTa (25%).

Mild ADEs expressed in positive sentiment as in the *Positive Sentiment* test do not pose a problem to the model. The performance on the *Positive Sentiment* test cases (72% for BioRedditBERT and 68% for XLM-RoBERTa) lies above the baseline of the positive class for both models. Also, the models' performance on the positive class negation test lies below the baseline, with BioRedditBERT (60%) again performing much better than XLM-RoBERTa (41%).

Unlike for the negative class test, almost all test cases in the *Beneficial Effect* test on the positive class are correctly classified as ADE. The poor performance on the negative *Beneficial Effect* test and the outstanding performance on the positive class *Beneficial Effect* test leads to the conclusion that the model has not learned to distinguish between ADEs and beneficial effects. Both models classify 96% of Beneficial Effects test cases as ADE, even though half of the tests have a no ADE mention. Possible explanations for this behavior are that the number of beneficial effect samples in the custom dataset is low and/or that the model does not take the context into account that distinguishes an ADE from a beneficial effect.

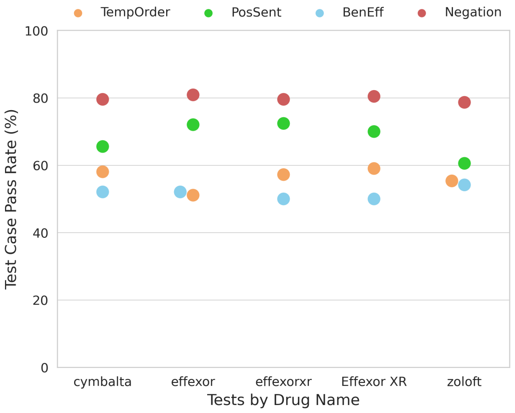

Figure 4: Performance of XLM-Roberta on test cases by drug name and by capability. The number of test cases per capability and drug name is 1440 (Temporal Order), 540 (Positive Sentiment), 48 (Beneficial Effect), 225 (Negation).

Each of the five selected drug name variants was used in every template allowing for an analysis of the impact of drug names in the test cases. Performance variations on test cases with different drug names indicate reduced robustness of the model. We find slight variations in the model performance over different drug names as shown in figure Figure 4 for XLM-Roberta. A potential explanation of these variations may be deviations in the occurrence of the respective drug names in the custom fine-tuning dataset, see Table 7 in Appendix B.

6 Conclusion and Future Work

In this work, we present a template-based approach for evaluating capabilities of models on the task of ADE detection in social media texts. Four capabilities, *Temporal Order*, *Positive Sentiment*, *Beneficial Effect*, and *Negation*, were identified and corresponding tests were created. Two high-performing models for the task of ADE detection were evaluated using the adapted approach.

Results show that the models' performances vary across capabilities. While both models perform well on the *Positive Sentiment* tests, BioReddit-BERT outperforms XLM-RoBERTa on *Negation*. The models are not able to distinguish between ADEs and beneficial effects and are not robust to changes in the expression of temporal structure in text. In summary, the template-based approach adapted to ADE classification has provided a better understanding of the shortcomings of high-performing models and can highlight previously undetected differences between models that perform almost identically on a held-out test set. We publish the templates to enable researchers to evaluate their own ADE classification models.

Further research may expand on this work by adding tests for more capabilities and evaluating other models using this approach. For example, in the phenomena annotation described in Section 3.2, we found 1.6% questions and 1.1% speculative content in the tweets. The linguistic variety of the templates could be improved by using a large language model to generate templates or test cases. A different direction of research may focus on improving the model's faults detected during evaluation. One method of improvement is to include a subset of the test cases as new training data (McCoy et al., 2019).

7 Limitations

While the approach of generating new inputs by templates undoubtedly has benefits, it also introduces some limitations. For instance, the combination of all entity fill-ins with all templates can produce some unnatural phrases. An example of this is the *Temporal Order* template "After taking {drug}, I had {ade}.". The ADE entity "weight gain" creates the unnatural sounding test case "After taking cymbalta, I had weight gain." instead of "After taking cymbalta, I gained weight." The unnatural use of language may introduce a bias. This should be kept in mind when using the templates. However, not all entity fill-ins will introduce such a bias and the model's performance on the test cases cannot be fully attributed to the effect of unnatural language use.

A second potential bias when using templates is that it may not be able to depict a large variety of language when only few templates were used. An example of this are the templates for the positive class *Beneficial Effect* test where each test case includes the word "problem". A model could use this as a proxy for correctly classifying the test cases.

Lastly, as described in Section 4, not all features of social media tests were used when creating templates. No anonymized usernames, hashtags, non-standard punctuation, and colloquialisms other than contractions were applied in the templates. This may introduce a bias as there is a slight difference in language variety between the templates and the training data. A researcher should keep in mind that slight changes in the model performance may also be attributed to this shift in language variety.

Acknowledgments

We would like to thank the anonymous reviewers for their feedback on the paper. We further thank our annotators Alon Drobickij, Selin Yeginer, and Emiliano Valdes Menchaca.

This work was partially supported by the German Federal Ministry of Education and Research as part of the project TRAILS. We are furthermore grateful for the support of the German Research Foundation (DFG) under the trilateral ANR-DFG-JST AI Research project KEEPHA (DFG-442445488), and the German Federal Ministry of Education and Research under the grant BI-FOLD24B.

References

Hiba Ahsan, Emmie Ohnuki, Avijit Mitra, and Hong You. 2021. Mimic-sbdh: A dataset for social and behavioral determinants of health. In *Machine Learning in Health Care*.

Hassan Alhuzali and Sophia Ananiadou. 2019. Improving classification of adverse drug reactions through using sentiment analysis and transfer learning. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 339–347, Florence, Italy. Association for Computational Linguistics.

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

Yonatan Belinkov and Yonatan Bisk. 2019. Synthetic and natural noise both break neural machine translation. *Conference paper at ICLR 2018*.

Shaily Bhatt, Rahul Jain, Sandipan Dandapat, and Sunayana Sitaram. 2021. A case study of efficacy and challenges in practical human-in-loop evaluation of NLP systems using checklist. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 120–130, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Oscar M. Cumbicus-Pineda, Itziar Gonzalez-Dios, and Aitor Soroa. 2021. Linguistic capabilities for a checklist-based evaluation in automatic text simplification. In *Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021) co-located with the 37th Conference of the Spanish Society for Natural Language Processing (SEPLN2021) Online (initially located in Málaga, Spain)*, pages 70–83.

Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, and Apurv Patki. 2014. Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark. In *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, pages 1–8.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892.

Andrey Gusev, Anna Kuznetsova, Anna Polyanskaya, and Egor Yatsishin. 2020. BERT implementation for detecting adverse drug effects mentions in Russian. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 46–50, Barcelona, Spain (Online). Association for Computational Linguistics.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An analysis of negation in natural language understanding corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 716–723, Dublin, Ireland. Association for Computational Linguistics.

Trung Huynh, Yulan He, Alistair Willis, and Stefan Rueger. 2016. Adverse Drug Reaction Classification With Deep Neural Networks. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 877–887.

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

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards Internet-age pharmacovigilance: Extracting adverse drug reactions from user posts in health-related social networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 117–125, Uppsala, Sweden. Association for Computational Linguistics.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv*, abs/1711.05101.

Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre-Maduell, Salvador Lima Lopez, Ivan Flores, Karen O'Connor, Davy Weissenbacher, Elena Tutubalina, Abeed Sarker, Juan M Banda, Martin Krallinger, and Graciela Gonzalez-Hernandez, editors. 2021a. *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*. Association for Computational Linguistics, Mexico City, Mexico.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021b. DeepADEMiner: A deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

Marta Marchiori Manerba and Sara Tonelli. 2021. Fine-grained fairness analysis of abusive language detection systems with CheckList. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 81–91, Online. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Zulfat Miftahutdinov, Andrey Sakhovskiy, and Elena Tutubalina. 2020. KFU NLP team at SMM4H 2020 tasks: Cross-lingual transfer learning with pretrained language models for drug reactions. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 51–56, Barcelona, Spain (Online). Association for Computational Linguistics.

Milad Moradi and Matthias Samwald. 2022. Improving the robustness and accuracy of biomedical language models through adversarial training. *Journal of Biomedical Informatics*, 132:104114.

Azadeh Nikfarjam and Graciela H. Gonzalez. 2011. Pattern mining for extraction of mentions of adverse drug reactions from user comments. In *AMIA annual symposium proceedings*, volume 2011, page 1019. American Medical Informatics Association.

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

Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting human-AI collaboration in auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 913–926. Association for Computing Machinery.

Marco Tulio Ribeiro and Scott Lundberg. 2022. Adaptive testing and debugging of NLP models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3253–3267. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. 2018. Data and systems for medication-related text classification and concept normalization from twitter: insights from the social media mining for health (smm4h)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

Simone Scaboro, Beatrice Portelli, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra. 2021. NADE: A benchmark for robust adverse drug events extraction in face of negations. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 230–237, Online. Association for Computational Linguistics.

Isabel Segura-Bedmar, Ricardo Revert, and Paloma Martínez. 2014. Detecting drugs and adverse events from Spanish social media streams. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 106–115, Gothenburg, Sweden. Association for Computational Linguistics.

Thinh Hung Truong, Yulia Otmakhova, Timothy Baldwin, Trevor Cohn, Jey Han Lau, and Karin Verspoor. 2022. Not another negation benchmark: The NaN-NLI test suite for sub-clausal negation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 883–894, Online only. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Davy Weissenbacher, Juan Banda, Vera Davydova, Darryl Estrada Zavala, Luis Gasco Sánchez, Yao Ge, Yuting Guo, Ari Klein, Martin Krallinger, Mathias Leddin, Arjun Magge, Raul Rodriguez-Esteban, Abeed

Sarker, Lucia Schmidt, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the seventh social media mining for health applications (#SMM4H) shared tasks at COLING 2022. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 221–241, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Tasks at ACL 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 21–30, Florence, Italy. Association for Computational Linguistics.

Yuqing Xie, Yi-An Lai, Yuanjun Xiong, Yi Zhang, and Stefano Soatto. 2021. Regression bugs are in your model! measuring, reducing and analyzing regressions in NLP model updates. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6589–6602, Online. Association for Computational Linguistics.

Maryam Zolnoori, Kin Wah Fung, Timothy B Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Yi Shuan Shirley Wu, Christina E Eldredge, Jake Luo, Mike Conway, et al. 2019. A systematic approach for developing a corpus of patient reported adverse drug events: a case study for ssri and snri medications. *Journal of biomedical informatics*, 90:103091.

A Templates

The number of templates with linguistic variations for each capability can be seen in Table 5. Example templates without filled-in entities are in Table 6.

capability	#base templates	#all templates
TempOrder	36	816
PosSent	36	504
BenEff	12	48
Negation	15	137
Total	**99**	**1505**

Table 5: Count of all created templates. Linguistic variation was used to create all templates from base templates.

A.1 Extraction of ADEs from PsyTAR

Sets of Parts of Speech combinations (tagsets) were created to define which sets of POS tags constitute a short noun phrase. An English POS tagger (spaCy) was then used to tag every token in the PsyTAR ADEs and filter out the chosen noun phrases. Examples of PsyTAR ADEs that were retrieved using this method are “listlessness”, “recurrence of ocular migraines”, and “bad pain in my right arm”. The goal of this process was to retrieve as many and diverse ADE descriptions as possible, yet the tagsets are not extensive and not all relevant ADEs were retrieved. Reasons for not passing the tagset filters were not being a noun phrase (“gained 18 pound”), incorrect POS tag assigned tagger (“heartburn”), incorrect POS tags assigned due to typos or extra whitespace, long noun phrases (“stomach cramping the first couple of days”), and punctuation marks/symbols (“increase in alcohol abuse/dependence”).

A.2 List of Beneficial Effects

List of (potential) beneficial effects used for the *Beneficial Effect* tests: weight loss/weight gain, sleepiness/decreased need for sleep, loss of appetite/increased appetite.

B Experiment Details

List of entities used as fill-ins for ADE, milder ADE for the *Positive Sentiment* test, and drug names used in the experiments for this project.

- drug names: zoloft, effexor, cymbalta, Effexor XR, effexorxr
- ADEs: Incredible sweet tooth, big appetite, many dreams, Difficulty Orgasming, excellerated heart rate, Insomnia, blackouts, bad pain in my right arm, a little more lethargy, VERY vivid dreams, stiff shoulders, EXTREME DRY MOUTH, Dialated pupils, increase in Libido, acid reflux
- milder ADEs: sugar craving, carbohydrate cravings, bouts of sleeplessness, gum pain, secretion under my toungue, weird dreams, stiff muscles, mild constipation, arm tingling, increased heat sensitivity, strange dreams, poorer concentration, cravings for sweets, hard time falling asleep, neck pain

The counts of the occurrences of the drug names can be found in Table 7.

C Model Details

BioRedditBERT (Basaldella et al., 2020) is a BERT-base uncased model related to BioBERT (Lee et al., 2019), a model pre-trained on the original BERT training corpus (English Wikipedia +

Test Name	Label	Test Description	Example Template
Temporal order standard	no ADE	ADE occurs before drug intake	Before taking {drug}, I experienced {ade}.
	ADE	ADE occurs after drug intake	Before having {ade}, I was put on {drug}.
Temporal order single time entity	no ADE	ADE occurs before drug intake expressed by a time entity	I was experiencing {ade} for {time_entity}, now I started being medicated with {drug}.
	ADE	ADE occurs after drug intake expressed by a time entity	{time_entity} ago I started being treated with {drug}, now I started encountering {ade}.
Temporal order double time entities	no ADE	ADE occurs before drug intake expressed by two related time entities	{time_entity_large} ago I started suffering from {ade}, I have been taking {drug} for {time_entity_small}.
	ADE	ADE occurs after drug intake expressed by two related time entities	I was enduring {ade} for {time_entity_small}, {time_entity_large} ago I started taking {drug}.
Positive Sentiment	ADE	ADE occurrence is reported with positive sentiment	I'm taking {drug} and experiencing {ade}. Still, I am happy my symptoms have reduced.
Beneficial Effect	no ADE	Secondary effect of a drug that is beneficial to the patient	I'm taking {drug} and experiencing weight loss. I'm happy because I was trying to lose weight anyway.
	ADE	Secondary effect of a drug that is an ADE as it is not beneficial	For me, weight loss is a side-effect of {drug}. It's a problem because I am already underweight.
Negation	no ADE	ADE is negated	I am taking {drug} without suffering from {ade}.
	ADE	Statement contains negation, ADE is not negated	That's not true, I took {drug} and encountered {ade}.

Table 6: Overview of all CheckList tests conducted for this project with example templates. Curly brackets in the example templates indicate entity placeholders.

	exact matches	all matches
cymbalta	451	742
effexor	172	312
effexorxr	0	0
Effexor XR	13	23
zoloft	50	100

Table 7: Occurrence of drug names in the fine-tuning training data. Exact matches are case-sensitive. A sample can contain multiple drug name occurrences. ”effexorxr“ was used in the templates without appearing in the training data.

BookCorpus) as well as on medical texts sourced from PubMed and PMC. BioRedditBERT, in turn, was initialized from BioBERT and continued to pre-train on a corpus of health-related Reddit posts. The Reddit dataset contains 800.000 posts from 68 health-related subreddits collected between 2015 and 2018. The specific set of training data of BioRedditBERT was the pivotal argument for choosing this model for the task of ADE classification on the Twitter dataset.

XLM-RoBERTa (Conneau et al., 2020) XLM-RoBERTa is a popular multilingual classification model without a focus on the biomedical domain.

We conducted hyperparameter search for both models and tried batch sizes of 8, 16 and 32 and learning rates of $3 \cdot 10^{-6}$, 10^{-5} and $3 \cdot 10^{-5}$. Both models achieved the best performance on the development set at $16, 10^{-5}$ and trained with the AdamW (Loshchilov and Hutter, 2017) optimizer. No truncation of inputs was applied and the model was evaluated on the validation set after every epoch. The inputs were sampled (batch sampling) with replacement weighted by class ratio due to the class imbalance (see Section 4.1.1).

D Per-Template Performance

The performance of the models on the template-based tests also varies within each test. For all tests except the *Beneficial Effect* tests, the models' performance varies for each template, see Figures 5 and 6. The dependence of the model performance on the template demonstrates that the wording of a template influences the models' ability to handle a capability. In turn, this stresses the importance of creating a wide range of variations in templates when using template-based evaluation.

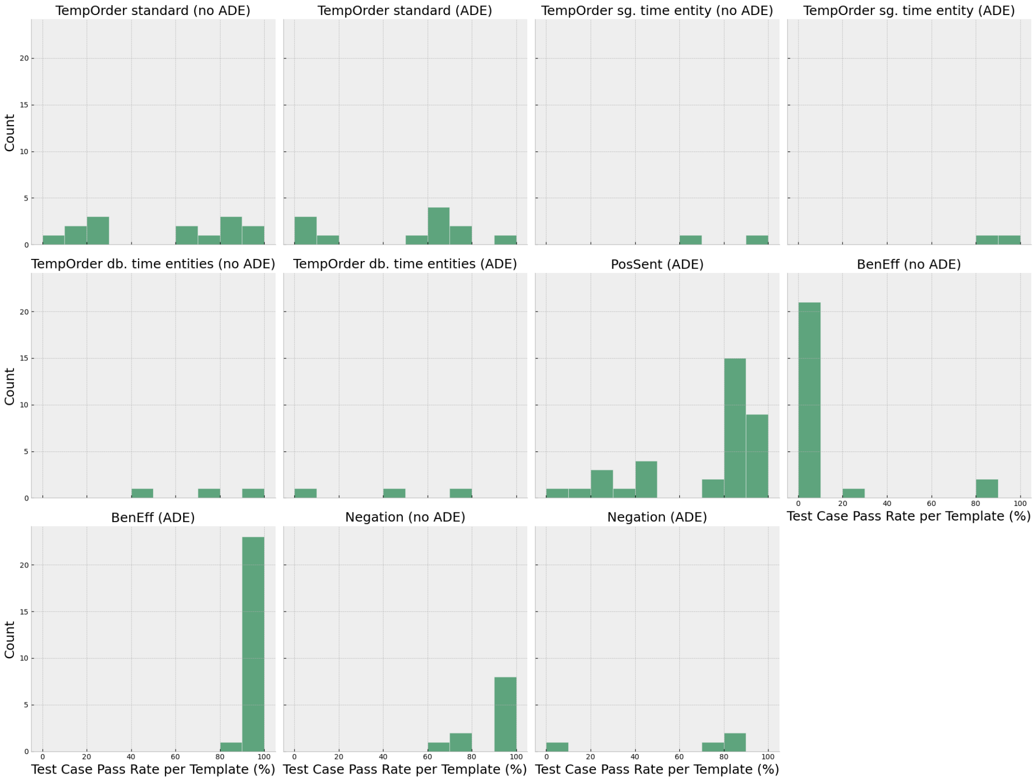

Figure 5: Results of the CheckList tests on the fine-tuned BioRedditBERT by template. The ratio of correctly classified test cases per template is shown on the horizontal axis. Each plot is a histogram showing the count of templates that produced more or less successfully classified test cases.

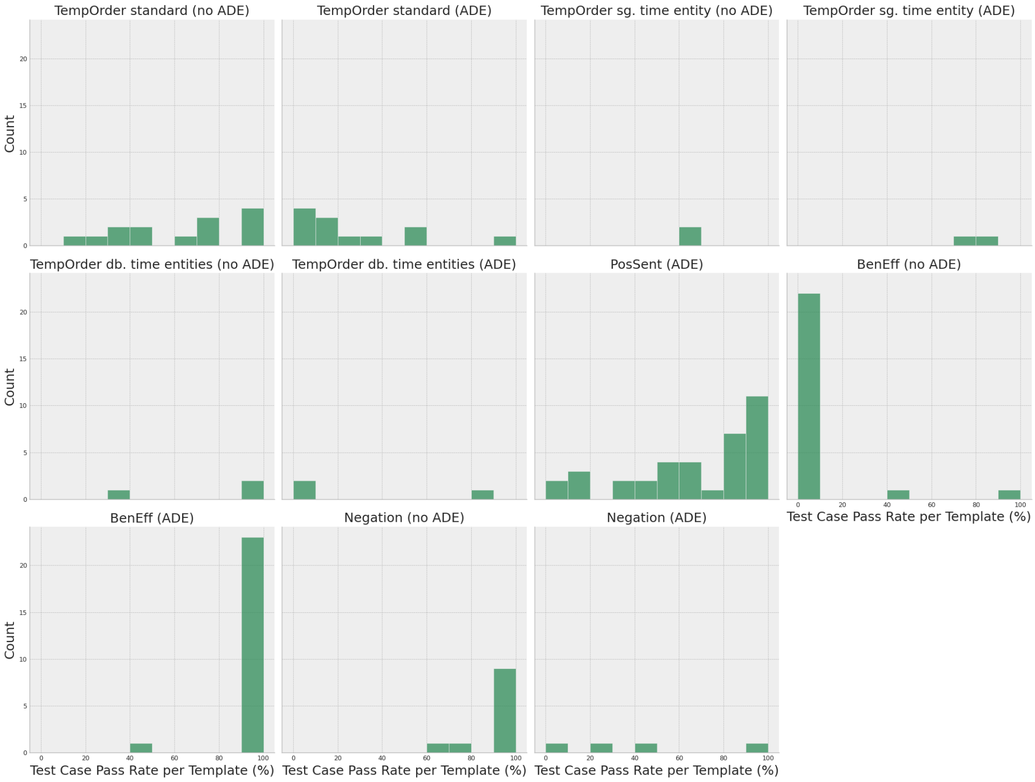

Figure 6: Results of the CheckList tests on the fine-tuned XLM-RoBERTa by template. The ratio of correctly classified test cases per template is shown on the horizontal axis. Each plot is a histogram showing the count of templates that produced more or less successfully classified test cases.