Domain- and Task-Adaptation for VaccinChatNL, a Dutch COVID-19 FAQ Answering Corpus and Classification Model

Jeska Buhmann, Maxime De Bruyn, Ehsan Lotfi, and **Walter Daelemans**
CLiPS Research Center
University of Antwerp, Belgium
{jeska.buhmann,maxime.debruyn,ehsan.lotfi,walter.daelemans}@uantwerpen.be

Abstract

FAQs are important resources to find information. However, especially if a FAQ concerns many question-answer pairs, it can be a difficult and time-consuming job to find the answer you are looking for. A FAQ chatbot can ease this process by automatically retrieving the relevant answer to a user's question. We present VaccinChatNL, a Dutch FAQ corpus on the topic of COVID-19 vaccination. Starting with 50 question-answer pairs we built VaccinChat, a FAQ chatbot, which we used to gather more user questions that were also annotated with the appropriate or new answer classes. This iterative process of gathering user questions, annotating them, and retraining the model with the increased data set led to a corpus that now contains 12,883 user questions divided over 181 answers. We provide the first publicly available Dutch FAQ answering data set of this size with large groups of semantically equivalent human-paraphrased questions. Furthermore, our study shows that before fine-tuning a classifier, continued pre-training of Dutch language models with task- and/or domain-specific data improves classification results. In addition, we show that large groups of semantically similar questions are important for obtaining well-performing intent classification models.

1 Introduction

In quickly changing contexts, like the COVID-19 pandemic, getting access to relevant and correct information in a fast and easy way is of crucial importance. Although websites with frequently-asked-questions (FAQ) sections provide such information, finding the representative question that matches a user's request can be hard and time-consuming, especially when the list of question-answer (QA) pairs is long. A FAQ chatbot does this matching for the user, speeding up the process of finding an answer to the posed question: users speak or type their questions and the system retrieves the best matching answer available.

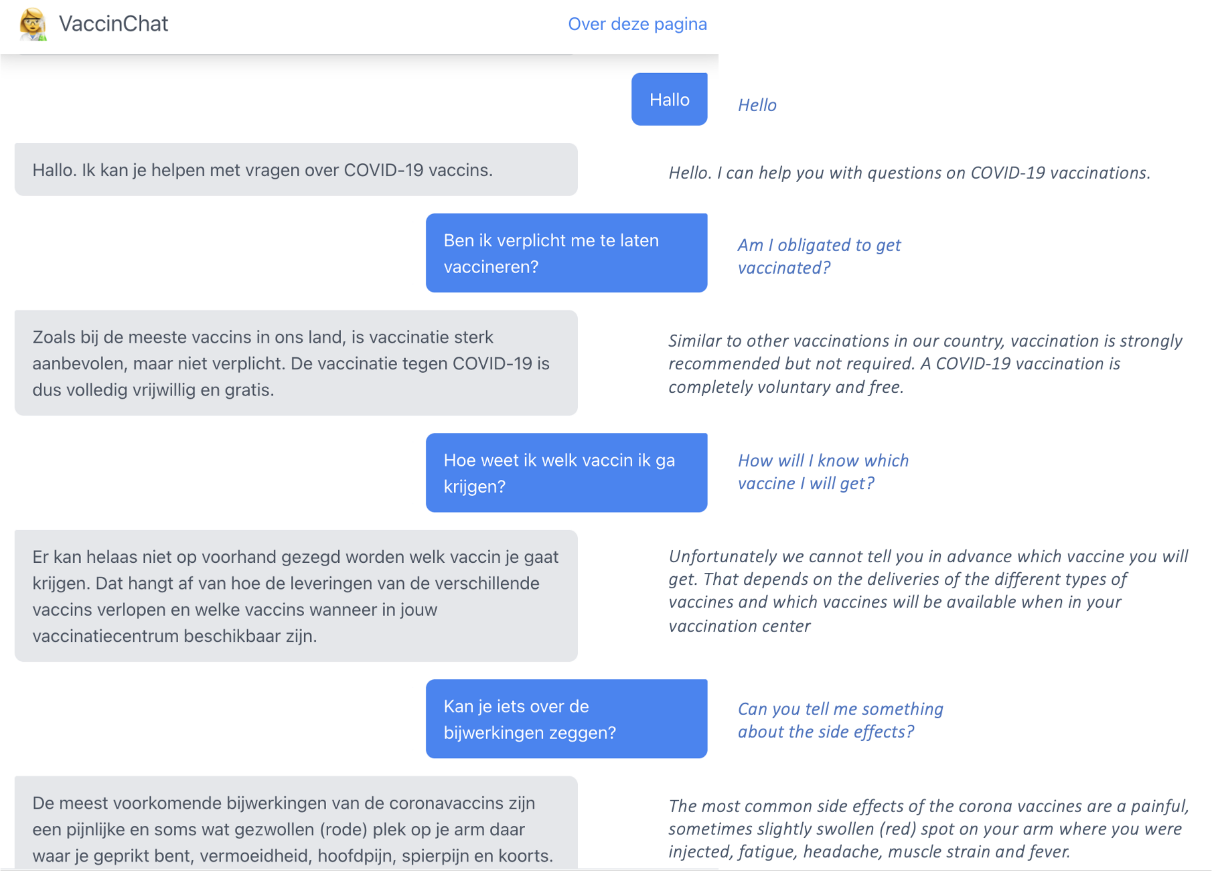
Intent: faq_ask_why
Waarom zou ik mij laten vaccineren? *Why would I get vaccinated?*
Wat zijn de voordelen? *What are the advantages?*
Waarom moet je je laten inenten? *Why do you need to get vaccinated?*
Ik snap niet waarom ik me moet laten vaccineren. *I don't understand why I must get vaccinated.*
Waarom is vaccineren belangrijk? *Why is vaccination important?*
Hoezo moet ik mij laten inenten tegen covid? *Why should I get vaccinated against covid?*
Waarom dringen ze zo aan op het vaccin? *Why are they so insistent on the vaccine?*
Intent: faq_ask_certificate
Krijgen we een attest na vaccinatie? *Do we get a certificate after vaccination?*
Moet ik later kunnen bevestigen dat ik een vaccinatie gehad heb? *Do I have to be able to confirm later that I have had a vaccination?*
Kunnen instanties me vragen te bewijzen of ik gevaccineerd ben? *Can authorities ask me to prove that I have been vaccinated?*

Table 1: Examples of paraphrases for two intents in the corpus.

Standard FAQs consist of a list of QA pairs, where each answer is coupled with one instance of a general question. However, to train an automatic classifier of answers, more than just one example of questions is needed. Different paraphrases or versions of the same question will improve the classification performance. In this work we describe the data collection process for VaccinChat[1],

[1] https://vaccinchat.be

Proceedings of the 29th International Conference on Computational Linguistics, pages 3539–3549
October 12–17, 2022.



Hallo

Hello

Hallo. Ik kan je helpen met vragen over COVID-19 vaccins.

Hello. I can help you with questions on COVID-19 vaccinations.

Ben ik verplicht me te laten vaccineren?

Am I obligated to get vaccinated?

Zoals bij de meeste vaccins in ons land, is vaccinatie sterk aanbevolen, maar niet verplicht. De vaccinatie tegen COVID-19 is dus volledig vrijwillig en gratis.

Similar to other vaccinations in our country, vaccination is strongly recommended but not required. A COVID-19 vaccination is completely voluntary and free.

Hoe weet ik welk vaccin ik ga krijgen?

How will I know which vaccine I will get?

Er kan helaas niet op voorhand gezegd worden welk vaccin je gaat krijgen. Dat hangt af van hoe de leveringen van de verschillende vaccins verlopen en welke vaccins wanneer in jouw vaccinatiecentrum beschikbaar zijn.

Unfortunately we cannot tell you in advance which vaccine you will get. That depends on the deliveries of the different types of vaccines and which vaccines will be available when in your vaccination center

Kan je iets over de bijwerkingen zeggen?

Can you tell me something about the side effects?

De meest voorkomende bijwerkingen van de coronavaccins zijn een pijnlijke en soms wat gezwollen (rode) plek op je arm daar waar je geprikt bent, vermoeidheid, hoofdpijn, spierpijn en koorts.

The most common side effects of the corona vaccines are a painful, sometimes slightly swollen (red) spot on your arm where you were injected, fatigue, headache, muscle strain and fever.

Figure 1: Example dialogue in VaccinChat.

a Dutch FAQ chatbot on the COVID-19 vaccines. The result is VaccinChatNL, the first Dutch FAQ answering data set, containing 12,883 questions that refer to 181 different answers. With Dutch we refer to the Dutch used in the Flemish region of Belgium as opposed to the Dutch used in the Netherlands. Table 1 shows some examples of user questions for two different answers and Figure 1 shows an example dialogue in VaccinChat.

Commercial organizations that use or develop FAQ chatbots, usually keep their FAQ answering data sets private. However, VaccinChat was built for research purposes and the released corpus was checked and anonymized with respect to privacy-sensitive information. The number of FAQ answers (181) is a lot higher than on the average FAQ page on the web holding only six answers (De Bruyn et al., 2021b). In addition, the corpus does not just hold QA pairs, but groups of questions per answer, providing a unique set of many-to-one mappings of questions to answers. Unlike FAQs on the web with 'clean', grammatically correct questions, the questions in our corpus come from actual users and may contain typing errors and/or other omissions/mistakes.

Questions referring to the same answer are not always paraphrases (e.g., "Should we get a vaccine on a yearly basis?" and "How long will the vaccine protect me?": both refer to the same answer), but in a lot of cases they are (see the examples in Table 1). So, although VaccinChatNL is not a paraphrase corpus in the strictest sense of the word, it does contain a lot of paraphrases. And unlike most of the available paraphrase corpora that contain paraphrase pairs, VaccinChatNL contains paraphrase groups that have many more than two paraphrases.

Our FAQ chatbot uses an intent classifier, meaning that user questions are classified into an answer class (intent), after which the answer of the classified intent is returned to the user. To train such a classifier we need the user questions and their intent labels. Since we also have the text of the answers, the VaccinChatNL corpus could be used to train a model that considers the similarity between (the representations of) a question and its answer. ConveRT (Henderson et al., 2019) is an example of such a model for English, and a Dutch version was developed by De Bruyn et al. (2021a).

On top of the above-mentioned qualities of the corpus, the topic of the corpus – COVID-19 vaccines – makes VaccinChatNL relevant. Although the data provides a blueprint of questions and answers for the Flemish situation in the first half of 2021, the topic is relevant and similar for the whole world. Translated data could be used for training models in other languages. Since Dutch (FAQ) data is scarce, the corpus is an excellent addition to multilingual FAQ approaches and data sets.

Despite the multipurpose nature of the corpus, we focus on using the data for training an intent classifier as one of the possible applications for the COVID-19 FAQ domain. Rather than finding the best possible model for this purpose, we focus on inspecting the effect of pre-training with task- and/or domain-specific data. In this work we describe those experiments as well as the data collection process. The VaccinChatNL corpus is publicly available[2] on the HuggingFace dataset hub (Lhoest et al., 2021).

2 Related Work

This section contains related work concerning domain- and task-adaptation, and considering the overlap of VaccinChatNL with FAQ and paraphrase corpora, we review the existing literature on these types of data resources.

2.1 Domain- and Task-Adaptation

Since large labeled data sets for classification are not always easy to build, pre-trained language models are used as a starting point for fine-tuning on the classification task. Edwards et al. (2020) show the importance of using domain-specific data for further unlabeled pre-training of a language model, improving performance for sentiment and emoji classification (among other things) of twitter utterances. Similarly, Wiedemann et al. (2020) show the benefit of domain-adaptation of a RoBERTa-large model for binary offensive language detection.

Zhu et al. (2021) found that domain-adaptation is mainly beneficial in low-resource settings. Increasing the amount of labeled data available for fine-tuning, reduces the impact of further pre-training with domain-specific data. Similar to the work of Mehri et al. (2020) they also investigated pre-training with task-specific data (i.e., data for the downstream task) and concluded as well that the impact of both domain- and task-adaptation depends on the type of task, the model, and the amount of data for fine-tuning. We therefore tested the impact of domain- and task-adaptation for our particular case, i.e., the size of our training set in combination with the BERT- (Devlin et al., 2019) and RoBERTa-based (Liu et al., 2019) text classification models we used (see section 4.2).

In this paper we not only analyse the impact of pre-training with task- and/or domain-specific data on the overall classification performance, but we also investigate the linguistic effects by doing an error analysis of how different intent types (FAQ, chitchat, and out-of-domain data) are affected.

2.2 FAQ and Paraphrase Data Sets

Faq-Finder (Hammond et al., 1995) created FAQs on multiple topics based on an English data set collected from Usenet news groups. The FAQIR dataset (Karan and Šnajder, 2016) contains 1,233 English queries retrieved from the “maintenance & repairs” section of the website *Yahoo! Answers*. A data set of similar size is StackFAQ (Karan and Šnajder, 2018), containing 1,249 user queries from the *StackExchange* domain. LocalGov is a Japanese FAQ corpus with 784 user queries in the domain of government (Sakata et al., 2019).

In the domain of question answering the semantic similarity between questions is relevant. The Kaggle Quora Question Pairs corpus[3] contains over 400,000 English question pairs labeled for semantic similarity.

Marsi and Krahmer (2014) developed a large Dutch aligned treebank corpus to study semantic similarity. It covers various text genres, such as texts from books, auto-cue subtitle pairs, news headlines, press releases about news events, and a section with QA-system output. This last part concerns a QA-system in the medical domain (van den Bosch and Bouma, 2011), that searched answers to questions in e.g., medical encyclopedia and layman websites. It resembles a FAQ answering data set, but instead of just QA pairs, it contains 100 questions that on average each have two answers (almost 200 answers in total). Although the complete corpus is large and consists of over 2M tokens, the size of the QA section is small and does not include variations or paraphrases of the questions.

The Dutch part of the multilingual TaPaCo

[2] https://huggingface.co/datasets/clips/VaccinChatNL

[3] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

corpus (Scherrer, 2020) contains sentential paraphrases: 23,561 paraphrases divided over 9,441 semantically similar paraphrase sets. This means that a group of similar semantics on average has 2.5 paraphrases. Because we collected user queries via our VaccinChat chatbot, the average number of queries per answer (including paraphrases) is a lot larger.

More specifically in the domain of COVID-19 Zhang et al. (2020) recently released COUGH, a multilingual FAQ retrieval dataset. A total of ∼16k FAQ items were scraped from 55 authoritative institutional websites, covering a wide range of topics on COVID-19, from general virus information to specific COVID-related instructions for a healthy diet. Almost 7k FAQ items were scraped from non-English FAQ sources. Besides FAQ items (QA pairs), the corpus also contains a Query Bank with 1,236 human-paraphrased user queries (three human paraphrases per given FAQ question), be it only for the English part of the corpus. Our VaccinChatNL corpus is similar to COUGH in the sense that both are FAQ data sets and the topic is similar, although COUGH covers a wider range of COVID-19 issues than VaccinChatNL, that narrows the focus on vaccines and vaccination. Unfortunately, the Dutch part of COUGH only consists of 142 QA pairs and no human-paraphrased versions of questions. VaccinChatNL by comparison consists of 12,883 questions referring to 181 different answers, so each answer on average has 71 different versions of questions referring to it. Also note that in VaccinChatNL the questions are actual user questions and not paraphrased versions of a set of given questions (as is the case for the English part of COUGH).

3 Dutch VaccinChatNL Corpus

In this section, we present our new Dutch VaccinChatNL corpus, a FAQ answering data set with many-to-one mappings of user questions to FAQ answers.

3.1 RASA Chatbot as a Data Collection Tool

Building a chatbot in RASA (Bocklisch et al., 2017) is an iterative, conversation driven, development process. This implies that the chatbot is continuously improved by (semi-automatically) annotating new user conversations or by adding new elements not yet present in the chatbot, after which the model is retrained. Because of this cyclic, user-driven way of development and an easy user interface, we chose RASA for building our chatbot, while simultaneously collecting user data.

3.2 Data Set Collection

3.2.1 Cold Start

We started with collecting approximately fifty FAQ pairs from the Belgian governmental webpage on vaccination[4]. Some of the answers were mildly adapted to better suit the conversational domain, i.e., sometimes answers were shortened to get the message across faster and we made sure none of the answers started with “yes” or “no”, making them applicable not only to yes/no questions but also to semantically similar questions or utterances.

Together with a fallback option, a start message, and a few basic chitchat intents (e.g., *chitchat_ask_bye* and *chitchat_ask_thanks*), these FAQ pairs were entered in a RASA environment, and a first version of our FAQ chatbot was trained.

3.2.2 Improvement Phase

In the second data collection phase we grew the corpus by having a small group of users test the chatbot. The user questions were annotated with the relevant answer classes. In addition, we performed error analysis on a 20% held-out data set to find out which user questions could benefit from adding more paraphrases. This iterative process also revealed missing information, leading to updates of already existing answers and the addition of new answers. In May 2021 VaccinChat was opened to the public generating a surge of user questions that had to be annotated. The growth of the corpus was maintained until July 2021, resulting in 181 different answers and 11,650 user questions in total.

3.2.3 Testing Phase

In addition to the above described data collection effort, another 1,267 user queries were collected and annotated for an ongoing study on the comparison of different versions of the chatbot (Poels et al., 2021). Participants in the study had no prior knowledge about the chatbot and were instructed to use the chatbot for five minutes by typing in questions they might have on COVID-19 vaccination. Before and after the chatting phase, they were given questionnaires about vaccination-related topics. The post-questionnaire also informed about the chatbot’s user-interface aspects.

[4] https://www.info-coronavirus.be/nl/vaccinatie

3.3 Data Annotation

As described in section 3.2, the data collection and manual annotation of the user queries happened incrementally. The complete VaccinChatNL data set was checked and corrected by three annotators (without overlap), starting from the RASA predictions. These annotators were computational linguists.

To assess the consistency in assigning the relevant answer classes and measure Inter-Annotator-Agreement, we, afterwards, had four different Dutch-speaking persons annotate a small subset (62 randomly selected user queries) of our complete annotated data set. Two of the annotators had a computational linguistics profile (one was not an author of the paper), and the other two had an end-user profile.

All four of the annotators were instructed to first get acquainted with the type of answers available by going over the list of possible intent names. They were also told to use the label *nlu_fallback* if the question was incomprehensible or completely irrelevant (e.g., an insult), and to use the label *faq_ask_general_information* if the question was COVID-19 related but not concerning vaccination. The raters were provided a document with all the intent names and answers, and were instructed to use this document to search for keywords related to the user query. Fleiss' kappa showed that there was good agreement between the raters' judgements, $\kappa = .663$.

3.4 Ethics and Privacy

Users of VaccinChat were informed that the chatbot could give irrelevant answers to their questions, and that their conversations would be used to further improve the performance of the chatbot.

Afterwards, the collection of data was checked for privacy-sensitive information, such as names and telephone numbers. Two user queries with names were removed, as well as three with telephone numbers. We also removed 28 entries containing codes that were used in the Testing phase (section 3.2.3) to link conversations to user questionnaires, and we removed three entries containing a URL.

Although we checked for other metadata that could reveal personal information (age, domicile, gender, disease, job), such entries were not deleted, because they could not be traced back to a specific person; in our corpus the combinations of metadata never revealed a specific identity, e.g, "persons aged 71 living in Antwerp" is not specific enough to characterize an individual person.

In addition, all annotated user queries are stored as separate QA pairs that no longer link back to their original user conversations. This means that information can not be combined with user information that was revealed somewhere else in that respective conversation.

3.5 Data Set Statistics

The final VaccinChatNL corpus consists of 12,883 user queries referring to 181 different answers. On average each answer class has 71 (*SD* = 92; *Mdn* = 47) user queries. The class with the highest number of user queries is the fallback class, also referred to as the unanswerable questions. The corpus contains 903 of such examples, representing 7.01% of the total number of user questions. The second biggest paraphrase group contains the user queries referring to the general side effects of the COVID-19 vaccines (*faq_ask_general_side_effects*). This group contains 416 paraphrases (3.23%).

3.6 Evolution over Time

During the COVID-19 pandemic the stream of information was constantly changing. When we started developing VaccinChat, vaccines were becoming available, and the vaccination campaign just began. One of the things people were asking questions about at that time was when they would be invited for vaccination and whether or not they belonged to a group that would get priority (answer class: *faq_ask_priority_groups*). A little later, people were most concerned with how long the vaccines would protect them against COVID-19 (answer class: *faq_ask_duration_of_protection*) and e.g., the blood clots side effects of the AstraZeneca vaccine (answer class: *faq_ask_astrazeneca_blood_clots*). Towards the summer of 2021, people started informing about their chances of going on holiday while also having to be available for their second vaccination (answer class: *faq_ask_holidays*)[5].

4 Experiments

In this section we describe a number of different classification models, built with VaccinChatNL data.

[5] A data visualisation of this evolution can be found at https://public.flourish.studio/visualisation/6517886/

	Train		**Dev**		**Test**	
Size (total 12,883)	10,542		1,171		1,170	
Intents	#	%	#	%	#	%
nlu_fallback	740	7.02	82	7.00	81	6.92
faq_ask_general_side_effects	303	2.87	56	4.78	57	4.87
faq_ask_priority_groups	291	2.76	46	3.93	45	3.85
faq_ask_protection_rate	225	2.13	34	2.90	35	2.99
faq_ask_general_information	200	1.90	55	4.70	55	4.70
faq_ask_no_risk_patient	200	1.90	18	1.54	18	1.54
faq_ask_astrazeneca_blood_clots	189	1.79	22	1.88	21	1.79
faq_ask_contra_indication	189	1.79	15	1.28	15	1.28
faq_ask_holidays	188	1.78	21	1.79	23	1.97
faq_ask_no_answer	174	1.65	36	3.07	35	2.99
faq_ask_duration_of_protection	153	1.45	24	2.05	24	2.05
faq_ask_trustworthy	137	1.30	27	2.31	28	2.39
faq_ask_choice	131	1.24	26	2.22	25	2.14
faq_ask_why	127	1.20	13	1.11	11	0.94
faq_ask_when_and_why	126	1.20	15	1.28	14	1.20

Table 2: Summary of VaccinChatNL data in terms of number of user questions per train, development and test set. The bottom part shows the absolute numbers (#) and percentages (%) of user questions for the 15 most frequent FAQ intents.

4.1 Train, Development, and Test Sets

To present the models' performance on unseen data, the corpus was split up in a train, a development, and a test set. The collected data in the Testing phase was mixed together with a random 10% of the data collected in the Improvement Phase. The merged data were split into a development and test set. The remaining 90% of the Improvement phase data was kept for training. Statistics of these sets are presented in Table 2, which also shows the absolute numbers and percentages of the fifteen most frequent intents in the corpus. The most frequent intent is the fallback option. This includes out-of-domain questions, insults, jokes, and questions that are incomprehensible because they contain either too many typing mistakes or only keywords that could refer to multiple answers.

4.2 Baseline Models

Here we present a number of baseline models for our VaccinChatNL corpus. We start with a simple majority class baseline and the RASA baseline classifier. In addition, we use pre-trained Dutch language models: BERTje (de Vries et al., 2019) and CoNTACT (Lemmens et al., 2022), based on BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), respectively.

4.2.1 Majority

The majority baseline always gives the same, most frequent answer. It thus always returns "*Ik begrijp het niet. Kunt u het anders zeggen?*" ("I don't understand. Could you rephrase the question?"). This is the fallback for unanswerable questions.

4.2.2 RASA DIET Classifier

RASA (Bocklisch et al., 2017) uses a multitask transformer architecture for NLU: Dual Intent and Entity Transformer (DIET)[6]. This DIET classifier can deal with both intent classification and entity recognition simultaneously. Only intent classification applies to our chatbot.

4.2.3 BERTje

BERTje is a Dutch pre-trained BERT model (Devlin et al., 2019) developed by de Vries et al. (2019). Several high-quality level Dutch corpora were used for pre-training, including collections of books, news articles, and Wikipedia documents.

4.2.4 BERTje+

This is the BERTje model, but further pre-trained with a masked-language learning (MLM) approach

[6] https://rasa.com/blog/introducing-dual-intent-and-entity-transformer-diet-state-of-the-art-performance-on-a-lightweight-architecture/

Number of train items		Test set		
Per class	Total	P	R	F1
1	181	0.01 (0.005)	0.01 (0.004)	0.01 (0.004)
2	351	0.02 (0.010)	0.02 (0.009)	0.02 (0.009)
4	679	0.07 (0.011)	0.07 (0.011)	0.07 (0.011)
8	1,331	0.20 (0.018)	0.19 (0.017)	0.20 (0.017)
10	1,655	0.23 (0.021)	0.23 (0.021)	0.23 (0.021)
30	4,648	0.41 (0.013)	0.41 (0.012)	0.41 (0.013)
100	8,591	0.63 (0.008)	0.63 (0.007)	0.63 (0.008)
300	10,099	0.65 (0.006)	0.65 (0.005)	0.65 (0.006)

Table 3: The effect of the training set size (number of train items per class) on test set performance: average Precision (P), Recall (R) and F1 from 10-fold cross-validations. Standard deviation is shown between brackets.

on task-specific data, i.e., the VaccinChatNL user queries excluding the *nlu_fallback* class. This model was used to show the effect of task adaptation on the classification task.

4.2.5 CoNTACT

CoNTACT[7] (Lemmens et al., 2022) is a Dutch RoBERTa-based language model, adapted to the domain of COVID-19 tweets by extra pre-training with this data of the Dutch RobBERT model (Delobelle et al., 2020). It was used to show the effect of domain adaptation on the classification task.

4.2.6 CoNTACT+

Similar to BERTje+, this is the CoNTACT model with extra MLM pre-training on the VaccinChatNL user queries. It shows the effect of domain- and task-adaptation simultaneously.

4.3 Train Settings

For the data set size experiments (see 5.1) the RASA DIET Classifier was trained for 12 epochs. This was based on initial experiments where training for more epochs revealed a maximum (accuracy) performance on the development set at 12 epochs.

For the domain- and task-adaptation experiments (see 5.2) hyper-parameter search revealed optimal results on the development set with a maximum sequence length of 64, a batch size of 16, learning rate of 2e-5, and warm-up ratio of 0.1. The BERTje models were trained for 7 epochs and a weight decay of 0.01, and the CoNTACT models for 10 epochs, with weight decay of 0.1. In all cases early stopping was applied based on the development set performance. This resulted in stopping after 7 epochs for BERTje, after 5 epochs for BERTje+, and after 6 epochs for CoNTACT and CoNTACT+.

5 Results

5.1 Effect of Data Set Size

To show the importance of having enough paraphrases per class to train a sentence classifier, we experimented with training the RASA baseline model (DIET classifier) with different numbers of training examples per class. Table 3 shows the average results (precision, recall, and F1) of 10-fold cross-validation experiments on the test set. Each model was trained for 12 epochs.

Note that the mentioned numbers of training examples per class only apply to the classes that have that many training examples. If a class has fewer items than the stated number, all training items for that class are used. The second column in the table shows the total number of training examples.

Results clearly show that in order to get a near-optimal performance on the test set, the number of training examples per class should at least be 100.

5.2 Effect of Models

Although the majority class (the fallback for unanswerable questions) has a high number of occurrences in the corpus (903), this class still only represents 7.00% and 6.92% of the data in the development and test set respectively. This means that the majority baseline model gives an accuracy of 6.9% on the test set.

Best results with the RASA DIET classifier on the development set were obtained with 12 epochs of training. We get a test set accuracy of 64.7%, which is a big improvement over the majority baseline.

[7] https://huggingface.co/clips/contact

Model	Test set accuracy
Majority	6.9
RASA DIET classifier	64.7
BERTje	74.7
BERTje+	77.7
CoNTACT	77.1
CoNTACT+	77.9

Table 4: VaccinChatNL accuracy scores (% correct) with different models. +: models with extra MLM pre-training on task-specific data, i.e., the VaccinChatNL user queries from the train set, excluding the out-of-domain ones labelled as *nlu_fallback*. Results for these models are better than for the models without task-specific pre-training. Overall, CoNTACT+, the model with domain- as well as task-adaptation, performs best.

All other models - BERTje(+) and CoNTACT(+) - were Dutch pre-trained language models fine-tuned on a sentence classification task with the VaccinChatNL train data set. The results of all these models are shown in Table 4. All BERTje and CoNTACT models are a clear improvement over the RASA DIET classifier, showing the importance of using language-specific (i.e., Dutch) pre-trained language models.

Table 5 provides more specifics on the misclassifications of the above models per answer type (fallback answers, chitchat, and FAQ answers). In general, all models with extra task- and/or domain-specific pre-training perform better than BERTje, which had no such continued pre-training. In the following section we show the effect of domain- and task-adaptation on the classification performance by means of an error analysis.

6 Discussion

When taking a more in-depth look into the type of errors made by the BERTje model and its domain- and task-adapted versions, a main observation is that all models make errors for user questions with infrequent words, like place names (e.g., "Kessel-Lo", "rotselaar") or typos (see Table 6).

With respect to the fallback and chitchat intents, the results show a better performance for the CoNTACT models than for the BERTje models (see Table 5). This suggests a benefit of pre-training with domain-specific data, in terms of a better recognition of the non-domain-specific user questions. Some examples of chitchat intents that are correctly classified with the CoNTACT models but misclassified with the BERTje models are shown in Table 7.

We also see a decrease in FAQ errors with domain-specific pre-training, but only for the case where no task-specific pre-training was done (CoNTACT vs. BERTje). For CoNTACT+, the model that was pre-trained on both domain- and task-specific data, we only see FAQ improvements compared to CoNTACT, but not compared to BERTje+. Unlike the FAQ intents, fallback and chitchat intents are better recognized with CoNTACT+ than with BERTje+, but there is no difference with CoNTACT. In general, CoNTACT+ shows the best performance.

For task-specific pre-training a clear improvement was observed for user questions containing domain terminology, i.e., words like *KU Leuven*, *Facebook*, *SARS-COV-2*, *Kovid 19*, *Jnj*, *QVAX*, etc. This was to be expected, since these words were actually in the task-specific data.

In summary, we can conclude that task-specific pre-training improves FAQ intent classification, whereas domain-adaptation mainly benefits the fallback and chitchat intent classification, i.e., the out-of-domain user questions.

7 Conclusions and Future Work

We have presented VaccinChatNL, the first Dutch FAQ answering corpus with over 12k of user queries about COVID-19 vaccines, and on average 71 example questions per answer. As the corpus holds lots of paraphrased versions of questions, it can serve as a paraphrase corpus as well. Existing Dutch paraphrase corpora, or Dutch parts of multilingual corpora are sometimes bigger, but do not contain sentential paraphrases (less long phrases or single words), or have less paraphrases per semantically similar group (typically paraphrase pairs). The many-to-one mapping characteristic of this corpus is unique in FAQ and paraphrasing corpora.

As an example application of VaccinChatNL, we described training an intent classifier and we have shown that for a classifier with many classes (i.e., different answers), performance improves with more examples of user questions per class.

This work also showed the importance of using domain- or task-adapted pre-trained language models for the fine-tuning task of sentence classification. More specifically, domain-adaptation (COVID-related tweets) led to improvements for the non-FAQ questions, whereas task-adaptation improved performance for FAQ questions with domain-specific terminology.

Nr. of errors	BERTje	BERTje+	CoNTACT	CoNTACT+
Total	299	263	271	261
Fallback intent	36	37	29	31
Chitchat intent	18	18	14	14
FAQ intent	245	208	228	216

Table 5: Number of errors for the test set with different models, split up per intent type: fallback, chitchat, and FAQ. The table shows the positive impact of task-specific pre-training (+models) on FAQ questions, and the benefit of domain-adaptation on mainly chitchat and fallback utterances: CoNTACT(+) vs. BERTje(+). Overall, CoNTACT+ is the best performing model.

User question	True label	Predicted label
Mag ik na het vaccin terug **reizne** *After the vaccine can I* ***travle*** *again*	faq_ask_holidays	faq_ask_what_after_vaccination
Moey ik mij laten vaccineren ***Doy*** *I have to get vaccinated*	faq_ask_obligatory	nlu_fallback
Hoe word ik **vrijwillgier**? *How do I become a* ***voluneter****?*	faq_ask_volunteer	faq_ask_why_and_when
worden gezonde mensen **zien** van corona *Do healthy people get* ***see*** *from corona*	faq_ask_why	nlu_fallback

Table 6: Dutch examples of misclassification made by all models: user questions with infrequent words like typos. Translated utterances in italics and typos in bold.

User question	True label	Predicted label (BERTje)
nou doeeiiii *Well byeee*	chitchat_ask_bye	nlu_fallback
DAg *BYe*	chitchat_ask_bye	nlu_fallback
ok ben niet neig overtuig maar dank u wel *okay not real convinced but thank you*	chitchat_ask_thanks	faq_ask_no_answer
Oké. Alles is duidelijk. *Okay. Everything is clear.*	chitchat_ask_thanks	chitchat_ask_bye

Table 7: Dutch examples of misclassification in the chitchat intent class made by the BERTje models, but not with the CoNTACT models. Translated utterances in italics.

Although our focus has been on classification, it would be interesting to use a retrieval approach for FAQ answer selection, because there is a certain amount of overlap in the answers. A predicted answer with a different label than what was annotated, is not necessarily wrong. It could still include information that is relevant to the user. In addition, the VaccinChatNL corpus would be a good source for e.g., the purpose of training generative paraphrase models. Instead of just sentence pairs, the corpus provides many examples that have similar semantic meaning relevant for the corresponding answers.

Acknowledgements

We thank the reviewers for providing constructive feedback. In addition, we thank the involved annotators for contributing to this work. This research received funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme.

References

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. *Computing Research Repository*, arXiv:1712.05181. Version 2.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2021a. ConveRT for

FAQ answering. *Computing Research Repository*, arXiv:2108.00719. Version 3.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2021b. MFAQ: a multilingual FAQ dataset. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 1–13, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *Computing Research Repository*, arXiv:1912.09582.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based language model. *Computing Research Repository*, arXiv:2001.06286. Version 2.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aleksandra Edwards, Jose Camacho-Collados, Hélène De Ribaupierre, and Alun Preece. 2020. Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5522–5529, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kristian Hammond, Robin Burke, Charles Martin, and Steven Lytinen. 1995. Faq finder: a case-based approach to knowledge navigation. In *Proceedings the 11th Conference on Artificial Intelligence for Applications*, pages 80–86. IEEE.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2019. ConveRT: Efficient and accurate conversational representations from transformers. *Computing Research Repository*, arXiv:1911.03688. Version 2.

Mladen Karan and Jan Šnajder. 2016. Faqir – a frequently asked questions retrieval test collection. In *Text, Speech, and Dialogue*, pages 74–81, Cham. Springer International Publishing.

Mladen Karan and Jan Šnajder. 2018. Paraphrase-focused learning to rank for domain-specific frequently asked questions retrieval. *Expert Systems with Applications*, 91:418–433.

Jens Lemmens, Jens Van Nooten, Tim Kreutz, and Walter Daelemans. 2022. CoNTACT: A Dutch COVID-19 adapted BERT for vaccine hesitancy and argumentation detection. *Computing Research Repository*, arXiv:2203.07362.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. *Computing Research Repository*, arXiv:1907.11692.

Erwin Marsi and Emiel Krahmer. 2014. Construction of an aligned monolingual treebank for studying semantic similarity. *Language Resources and Evaluation*, 48(2):279–306.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. DialoGLUE: A natural language understanding benchmark for task-oriented dialogue. *Computing Research Repository*, arXiv:2009.13570. Version 2.

Karolien Poels, Toni Claessens, Jeska Buhmann, Maxime De Bruyn, Heidi Vandebosch, Pierre Van Damme, and Walter Daelemans. 2021. A chatbot as an intervention to foster public engagement with the covid-19 vaccines. In *Book of Abstracts of European Conference on Health Communication*, pages 89–90.

Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. Faq retrieval using query-question similarity and bert-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 1113–1116, New York, NY, USA. Association for Computing Machinery.

Yves Scherrer. 2020. TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages . In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages OP –. European Language Resources Association (ELRA).

Antal van den Bosch and Gosse Bouma. 2011. *Interactive multi-modal question-answering*. Springer Science & Business Media.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. 2020. UHH-LT at SemEval-2020 task 12:

Fine-tuning of pre-trained transformer networks for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1638–1644, Barcelona (online). International Committee for Computational Linguistics.

Xinliang Frederick Zhang, Heming Sun, Xiang Yue, Simon Lin, and Huan Sun. 2020. COUGH: A challenge dataset and models for COVID-19 FAQ retrieval. *Computing Research Repository*, arXiv:2010.12800. Version 2.

Qi Zhu, Yuxian Gu, Lingxiao Luo, Bing Li, Cheng Li, Wei Peng, Minlie Huang, and Xiaoyan Zhu. 2021. When does further pre-training MLM help? an empirical study on task-oriented dialog pre-training. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 54–61, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.