CoNTACT:
A Dutch COVID-19 Adapted BERT for Vaccine Hesitancy and Argumentation Detection[*]

Jens Lemmens, Jens Van Nooten, Tim Kreutz, and Walter Daelemans

CLiPS (University of Antwerp)
Lange Winkelstraat 40, 2000 Antwerp
`firstname.lastname@uantwerpen.be`

Abstract. We present CoNTACT[1]: a Dutch language model adapted to the domain of COVID-19 tweets. The model was developed by continuing the pre-training phase of RobBERT [3] by using 2.8M Dutch COVID-19 related tweets posted in 2021. In order to test the performance of the model and compare it to RobBERT, the two models were tested on two tasks: (1) binary vaccine hesitancy detection and (2) detection of arguments for vaccine hesitancy. For both tasks, not only Twitter but also Facebook data was used to show cross-genre performance. In our experiments, CoNTACT showed statistically significant gains over RobBERT in all experiments for task 1. For task 2, we observed substantial improvements in virtually all classes in all experiments. An error analysis indicated that the domain adaptation yielded better representations of domain-specific terminology, causing CoNTACT to make more accurate classification decisions.

Keywords: BERT · domain adaptation · COVID-19 · vaccine hesitancy · argumentation detection · social media

1 Introduction

We present CoNTACT (Contextual Neural Transformer Adapted to COVID-19 Tweets). The model was developed by fine-tuning RobBERT [3] (a RoBERTa-base model [13] pre-trained on Dutch data) on masked language modeling using 2.8M Dutch-language tweets related to COVID-19 that were posted in 2021. The model was evaluated on two tasks: (1) binary vaccine hesitancy classification and (2) classification of arguments for vaccine hesitancy. In order to measure the effect of the domain adaptation, the results were compared to out-of-the-box RobBERT. Moreover, the aforementioned tasks were not only performed on tweets, but also on Facebook comments to show the cross-genre benefits of the domain adaptation. Afterwards, a qualitative error analysis was conducted

---

[1] The model is available at `https://huggingface.co/clips/contact`

to show where CoNTACT improved compared to RobBERT and where it could potentially improve further. In earlier research, an English language model pre-trained on COVID-19 related tweets (COVID-Twitter-BERT) was developed [16]. We apply the same methodology for the first time to Dutch and extensively test the effect on two COVID-19 related classification tasks.

2   Related research

Traditional machine learning assumes that models are trained and tested on large amounts of data from the same domain, which is not always feasible due to lack of labelled data. Transfer learning, which involves the transfer of knowledge from one domain to another, is a technique that has been utilized successfully in machine learning, both in NLP (e.g. [23], [5], [17]) and computer vision (e.g. [20], [18], [24]) to combat this issue. An effective approach to transfer learning used frequently in recent years is the pre-training of language models, such as BERT [4], on large amounts of unsupervised data. The knowledge from this pre-training phase is then transferred to the subsequent fine-tuning phase on task- and domain-specific data, which has shown significant improvements on several benchmark datasets, e.g., [12] and [1]. Subsequently, several language- and domain-specific adaptations of language models have been developed for non-English data or to further improve the performance of the original models on specific tasks. Examples are BERTje and RobBERT [19] [3], the Dutch equiv-alents of BERT and RoBERTa, respectively), and CamemBERT [14], a French BERT model.

Domain adaptation, a special case of transfer learning where the model is first trained on unsupervised data from the domain of an intended task, aims to improve results even further "by minimizing the difference between the domain distributions" ([6], p. 1), thus creating a model that optimally learns from the training data. Regarding the domain of COVID-19, COVID-Twitter-BERT [16], a BERT-large model pre-trained on COVID-19 tweets, has shown statistically significant gains over the baseline BERT-large in various applications, including vaccine stance classification.

In other research related to vaccine stance classification, various rule-based, statistical and deep learning approaches for the classification of stance towards vaccines have been compared [8]. Concretely, the task consisted of multiclass classification of vaccine stance in social media messages ("for", "againts" or "undecided"). The authors concluded that both pre-trained language models and statistical ensemble models achieved equally high results on this task. This work focused on vaccine stance in general, but since the start of the COVID-19 pandemic vaccine stance classification has become almost inextricably linked to COVID-19 due to its societal relevance. An example is [22] who present CoV-axLies, a COVID-19 vaccine misinformation dataset, and demonstrate that their model, based on knowledge graphs, outperforms widely used classification meth-ods for the detection of vaccine misinformation, an important cause of vaccine hesitancy.

Specifically for Dutch, [21] collected Dutch tweets using keywords, and comments from Reddit and Nu.nl[2] threads related to COVID-19 in order to investigate polarity ("positive"/"negative") and stance ("support"/"reject"/"other") towards face masks and the social distance measure between March and October 2020. For polarity analysis, the Pattern library [2] was used, whereas manual annotations were used to train a stance classifier consisting of a linear feed forward neural network using stochastic gradient descend and a subword embedding layer, which achieved a test set accuracy of 65%. After applying the polarity analyzer and stance classifier to the above-mentioned data, it was shown that a more negative polarity was found in COVID-19 related messages than in a subset of messages that were unrelated to COVID-19. More specifically, a more negative polarity (and also stance) was found in messages mentioning face masks than in messages mentioning the social distancing measure. The various social media platforms that were used showed similar trends over time.

3   Methodology

3.1   Domain adaptation

For the development of CoNTACT, we utilize RobBERT [3], a Dutch RoBERTa model with 12 attention layers and 12 heads with 117M parameters trained on the Dutch segment of the OSCAR corpus (6.6B words). In line with [7], we approached adapting RobBERT by continuing its pre-training phase, that is by performing masked language modeling. For this task, we scraped Dutch-language tweets posted in 2021 using the Twitter API and the keyword method described in [10]. Then, all tweets related to COVID-19 were filtered from this Twitter collection using regular expressions based on inflected forms, part-of-speech tag variations and spelling variations of the keywords shown in Table 1.

Afterwards, all duplicates and retweets were filtered from this subset of COVID-19 related tweets. To detect retweets, we based ourselves on the "retweet status" attribute returned by the Twitter API and searched for tweets beginning with "RT @". Finally, the FastText language detector was used to remove all tweets that were not written in Dutch [9]. In the end, 2.8M tweets (66.8M tokens, split by whitespace) remained for the domain adaptation, which were

---

[2] Nu.nl is a Dutch news website that allows visitors to comment on news articles

Table 1. Keyword lemmas used to construct regular expressions for collecting COVID-19 related tweets (translated from Dutch to English).

Key words
corona, COVID-19, SARS-CoV-2, virologist, virus
vaccine, vaccinate, Astrazeneca, Pfizer, Moderna, Johnson & Johnson, Curevac, Sputnik
mouth mask, social distancing, bubble, contact tracing, quarantine, lockdown, curfew, 1.5m, cuddle contact

anonymized by replacing all tokens starting with "@" by "@USER". In order to estimate the precision, 300 randomly selected tweets were manually read and it was determined whether they were Dutch and relevant to the domain of COVID-19. This manual evaluation shows that our keyword extraction method has a precision of 90.0%. False positives included messages about other viruses and vaccines, such as the flu/influenza, and a single tweet in Afrikaans that did not get detected by the language detector.

For the domain adaptation, the 2.8M tweets mentioned above were used to continue RobBERT's pre-training phase for 4 epochs, using the default learning rate and the largest batch size that fit working memory (32). A loss of 1.702 was achieved on a validation set consisting of 20% of our data.

3.2   Data and experiments

To determine the effect of the domain adaptation, i.e., whether CoNTACT performs significantly better than RobBERT on tasks involving social media data related to COVID-19, the models were tested on two classification tasks: (1) vaccine hesitancy detection and (2) detection of arguments for vaccine hesitancy. The corpus used for the classification tasks was first described in [11], it consists of approx. 8,800 tweets and 5,200 Facebook comments annotated for vaccine stance and argumentation. Regarding the stance, possible class labels were "anti-vaccination", "vaccine-hesitant", "neutral" and "pro-vaccination", but these were converted to binary labels: "anti-vaccination" and "vaccine-hesitant" comprise the "hesitant" category, whereas the "not hesitant" category consists of all "neutral" and "pro" comments. The annotation scheme for vaccine hesitancy arguments on the other hand consisted of the following labels:

1. **Development**: messages that express worry about the development, testing methodology, distribution and public access of vaccines.
2. **Liberty**: messages that express concerns about how vaccines and vaccine laws affect civil liberty and personal freedom.
3. **Institutional motives**: messages expressing mistrust in motives of political or economic entities involved with vaccines.
4. **Efficacy**: messages claiming that vaccines are not efficient (enough) or unnecessary.
5. **Safety**: messages that express worry towards the safety of the vaccines and their side effects.
6. **Criticism on the vaccination strategy**: messages criticizing the government's vaccination strategy/campaign.
7. **Alternative medicine**: messages that prefer other means of protection over vaccines.
8. **Conspiracy theories**: messages that spread conspiracy theories about vaccines.

Table 2. Vaccine hesitancy data used for the cross validation experiments.

Class	Twitter	Facebook	Total
hesitant	1250	1250	2500
non-hesitant	1250	1250	2500
Total	2500	2500	5000

Table 3. Vaccine hesitancy arguments data used for the cross validation experiments.

Class	Twitter	Facebook	Total
alternative medicine	175	56	175
conspiracy theory	687	228	915
criticism on vaccination strategy	979	1,222	2,201
development	565	511	1,076
efficacy	860	400	1,260
institutional motives	1,189	312	2,131
safety	1,493	1,416	2,909
none	1,153	298	1,451
n messages	8,439	3,917	12,356

For vaccine hesitancy detection, both RobBERT and CoNTACT were fine-tuned with 10-fold cross validation. These cross validation experiments were performed in same-genre settings (fine-tuning and testing on tweets only; fine-tuning and testing on Facebook comments only) and mixed-genre settings (fine-tuning and testing on both Facebook and Twitter). Additionally, cross-genre experiments were conducted by fine-tuning on all Twitter data and testing on all Facebook data (and vice versa) in order to show the usefulness of CoNTACT when no data from an intended platform is available for fine-tuning. In order to avoid overfitting on a certain class or platform due to unbalanced data, a subset that was balanced by class and social media platform was used. The statistics of this subset can be found in Table 2. For all experiments, the default batch size (8) and learning rate (5e-5) was used and fine-tuning was performed for 4 epochs.

For the argumentation detection task, 8,439 tweets and 3,917 Facebook comments were used (i.e. all of the available vaccine-hesitant messages). The distribution of the arguments varies across the two social media platforms, as can be derived from Table 3. Further, it should be noted that vaccine-hesitant entries without any clear argumentation were used as negative examples for the models to learn from. Similarly to the stance detection task, the aforementioned data was used to fine-tune both RobBERT and our CoNTACT model. For the same- and mixed genre experiments, cross validation was used, whereas a train-test split was used for the cross-genre experiments. Since the data is heavily unbalanced in terms of argument distribution, however, we chose to conduct experiments with 5-fold instead of 10-fold cross validation in order to preserve more entries per test set. For all experiments, the default batch size (8) and learning rate (5e-5) were used and fine-tuning was performed for 4 epochs.

Table 4. Results (%) for vaccine hesitancy detection, including standard deviations (if applicable). The results are reported on the positive class, and statistically significant gains over the baseline are indicated with asterisks.

Model	Fine-tune	Test	Pre	Rec	F1	*
RobBERT	Twitter	Twitter	76.1 (3.6)	74.2 (4.3)	75.1 (3.1)	N/A
	Twitter	Facebook	62.0 (-)	59.8 (-)	60.9 (-)	N/A
	Facebook	Facebook	69.5 (3.1)	57.2 (3.2)	62.7 (2.6)	N/A
	Facebook	Twitter	67.4 (-)	63.0 (-)	65.1 (-)	N/A
	Both	Twitter	77.1 (2.8)	73.9 (4.0)	75.4 (-)	N/A
	Both	Facebook	70.6 (3.5)	64.6 (3.7)	67.4 (2.7)	N/A
CoNTACT	Twitter	Twitter	77.2 (3.5)	76.9 (4.1)	77.1 (3.6)	*
	Twitter	Facebook	65.2 (-)	64.9 (-)	65.0 (-)	***
	Facebook	Facebook	71.2 (3.2)	67.5 (3.1)	69.3 (2.9)	***
	Facebook	Twitter	71.0 (-)	82.5 (-)	76.3 (-)	***
	Both	Twitter	78.9 (4.2)	77.4 (1.7)	78.1 (2.5)	**
	Both	Facebook	73.2 (3.0)	68.2 (4.3)	70.6 (2.6)	**

4   Results

4.1   Vaccine hesitancy detection

In Table 4, the results of the experiments for vaccine hesitancy detection can be found. For the same-genre and mixed-genre experiments, the provided results (precision, recall, F1-score) are the averages of the test set scores on the positive class (i.e. vaccine hesitancy) in each cross validation split (the standard deviations are mentioned between brackets). For the cross-genre experiments, on the other hand, results are reported on the test sets. In cases where CoNTACT outperformed RobBERT, p-values were calculated to determine whether the observed improvements are statistically significant [15].

As shown in the results, both models perform better on Twitter data than on Facebook data, and fine-tuning on both platforms simultaneously yields higher results than fine-tuning on the individual platforms. The standard deviations are, in spite of the small test sets, relatively small, which indicates consistent model performance. When comparing the results of RobBERT to those of CoNTACT, it can be observed that CoNTACT outperforms RobBERT in all experimental settings with statistical significance, including the cross-genre experiments. In other words, when fine-tuning on Twitter but testing on Facebook, CoNTACT strongly outperforms RobBERT, although no Facebook data was used during it's domain adaptation or fine-tuning phase. Additionally, CoNTACT outperforms RobBERT on Facebook data even if the former is fine-tuned on Twitter data and the latter is fine-tuned on Facebook data (i.e. data from the same platform). These results highlight the cross-genre potential of CoNTACT.

In order to gain insight into which specific improvements CoNTACT made, a manual analysis[3] of the instances where CoNTACT classified vaccine stance

---

[3] All examples provided below were translated from Dutch to English.

correctly, and RobBERT did not, was conducted (for all experiments). False negatives, i.e., the cases where RobBERT did not predict vaccine hesitancy, but CoNTACT did (correctly), were the largest group of errors. They were found in vaccine-hesitant instances referring to pro-vaccination opinions, such as "'We do not have evidence that vaccines cause damage to pregnant women so we advise pregnant women to get vaccinated', what kind of an idiot says things like this?!". Further, false negatives were caused by sarcasm and other forms of implicit language, e.g. "they should start [the vaccination campaign] in Den Hague... double dosis". This message seems to express pro-vaccination opinions on a superficial level, but the author actually hopes that the government (located in Den Hague) will suffer from major side effects of the vaccine.

In comparison, false positives, i.e., cases where RobBERT incorrectly detected vaccine hesitancy, but CoNTACT correctly did not detect vaccine hesitancy, were found in messages containing certain hashtags or terms that are associated with vaccine hesitancy. For example, the tweet "#vaccinationobligation, because infecting others is not a fundamental right", expresses a pro-vaccination opinion. RobBERT, however, incorrectly detected vaccine hesitancy in this tweet, presumably because of the hashtag "#vaccinationobligation", which occurs frequently in vaccine-hesitant messages. Especially in the cross-genre experiments where the models were fine-tuned on Facebook and tested on Twitter, RobBERT was frequently confused by vaccine related hashtags, causing both false positives and negatives, whereas CoNTACT showed more understanding of said hashtags, even when both pro- and anti-vaccination hashtags appeared in the same message. Other false positives by the baseline were found in cases where vaccine-hesitant opinions were quoted or referred to, such as "'poison vaccine', yeah right, you're so childish". Similarly, pro-vaccination messages expressing a negative sentiment towards, for example, vaccination policy, were misclassified more often by RobBERT than by CoNTACT, e.g. "I am #provaccination but I support protest against the mismanagement of the government".

An additional analysis of the comments where CoNTACT failed to correctly predict the stance but RobBERT did not was conducted. False negatives (the smallest group of errors) were found in messages using implicit or sarcastic language, such as "this press conference was very clear as always...", similarly to the false negatives found in RobBERT. Regarding the false positives, the largest group of errors, we observed that there were cases where specific terms used frequently in vaccine-hesitant messages caused confusion, as was also observed in the error analysis of RobBERT. For example, in "those #SideEffects are not as bad as people think" and "#vaccineobligation is a must", CoNTACT interprets the hashtags as indicators for vaccine hesitancy, because it has learned this during the fine-tuning period. In conclusion, we observed that the models have difficulties with the same types of comments: messages containing forms implicit language caused false negative errors, whereas domain-specific terminology caused false positive errors. CoNTACT, however, made significantly less errors in these challenging cases due to the domain adaptation, indicating that CoNTACT has improved representations of COVID-19 related terminology.

Table 5. Precision, recall, F1-score and EMR of RobBERT and CoNTACT for the vaccine argument experiments.

Model	Fine-tune	Test	Pre	Rec	F1	EMR
RobBERT	Twitter	Twitter	62.5 (0.8)	50.2 (1.4)	55.0 (1.0)	46.7 (6.1)
	Twitter	Facebook	50.7 (-)	29.7 (-)	36.3 (-)	24.2 (-)
	Facebook	Facebook	48.4 (1.4)	31.7 (1.8)	37.3 (1.7)	34.9 (1.2)
	Facebook	Twitter	59.5 (-)	30.0 (-)	33.3 (-)	33.8 (-)
	Both	Twitter	62.9 (1.5)	53.4 (0.5)	57.3 (0.8)	47.7 (0.8)
	Both	Facebook	56.6 (2.8)	43.9 (3.1)	48.9 (2.9)	39.3 (1.6)
CoNTACT	Twitter	Twitter	64.7 (1.5)	56.2 (0.9)	59.8 (0.9)	49.2 (1.3)
	Twitter	Facebook	56.9 (-)	36.1 (-)	42.7 (-)	26.9 (-)
	Facebook	Facebook	55.5 (5.9)	41.1 (1.2)	46.2 (1.9)	41.0 (1.1)
	Facebook	Twitter	57.5 (-)	39.4 (-)	41.4 (-)	34.5 (-)
	Both	Twitter	64.1 (1.3)	58.4 (1.6)	60.9 (0.9)	49.5 (1.1)
	Both	Facebook	60.1 (3.3)	49.7 (2.2)	53.9 (2.4)	41.9 (1.1)

4.2   Argument classification

In Table 5 the results on argument classification are summarised. Precision, recall, F1 (incl. standard deviations), and exact match ratio (EMR), an accuracy score for cases where the entire set of labels was predicted correctly, are reported. Overall, both models perform better on Twitter than on Facebook data, including the cross-genre experiments, similarly to the stance classification experiments. Although CoNTACT outperforms RobBERT in the cross-genre experiments, the results are still noticeably lower than the in-genre experiments. Further, fine-tuning on both Facebook and Twitter simultaneously increases model performance. When comparing the models, it can be observed that CoNTACT outperforms RobBERT in all experiments.

The results for the individual arguments for RobBERT and CoNTACT are presented in Table 6 and 7, respectively. The provided results are the results on the positive classes in the test set(s). Regarding the baseline results, it can be observed that certain classes are predicted substantially better than others. Overall, RobBERT predicted the "safety" and "liberty" classes best, whereas the most difficult classes were "development" and "alternative medicine" (these were also the most underrepresented classes in our data).

When comparing the results of RobBERT to those of CoNTACT, an increase in performance on all classes in all experiments can be observed, except for "conspiracy theory" in Twitter when fine-tuning on both platforms, "alternative medicine" in Twitter when fine-tuning on Facebook, and "institutional motives" in Facebook when fine-tuning on Twitter. Some of the highest improvements were found in the "development" and "alternative medicine" classes, which are the most challenging classes, as mentioned above. In order to verify whether the observed improvements are significant, a McNemar [15] test was conducted per argument class (Table 8). Despite the substantial gains, less than half of the improvements were considered statistically significant for the same- and mixed-genre experiments. We suspect that the significance test we used yielded higher

p-values because the frequency of certain classes was too low to ascertain that improvements were significant rather than random. Further experiments with more data could therefore produce other results and new insights in the future. In the cross-genre experiments, however, CoNTACT showed statistically significant improvements on half of the argumentation classes when the model was fine-tuned on Twitter data and tested on Facebook data. Moreover, statistically significant improvements were observed for all classes when the model was fine-tuned on Facebook data and tested on Twitter data. These results highlight the cross-genre potential of CoNTACT.

Table 6. Averaged results (%) of RobBERT on each argument class per experiment.

	tw-tw			tw-fb			fb-fb			fb-tw			both-tw			both-fb		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
alt.	60	35	45	33	36	35	0	0	0	100	6	11	66	46	54	45	32	38
con.	61	43	50	32	14	19	48	22	30	34	10	15	59	46	52	54	38	45
crit.	47	27	34	45	26	33	57	49	53	27	36	31	49	30	37	59	51	55
dev.	54	38	45	42	18	25	50	18	26	54	18	27	55	45	49	47	31	37
eff.	63	54	58	50	46	48	52	36	42	61	31	42	61	55	58	58	46	51
inst.	66	60	63	59	27	37	59	32	41	76	9	17	66	62	64	60	38	46
lib.	77	76	77	61	36	46	61	48	54	64	83	72	77	78	77	64	49	56
saf.	71	67	69	84	34	49	66	63	64	60	47	53	70	67	69	67	66	67
micro	69	60	64	58	30	39	59	44	50	57	45	50	68	62	65	61	51	56
macro	62	50	55	51	30	36	48	32	37	60	30	33	63	53	57	60	44	49
weighted	67	60	63	61	30	39	57	44	48	60	45	45	67	62	64	60	51	55
samples	56	53	53	32	28	29	50	45	46	50	42	44	58	55	55	52	49	49

Table 7. Averaged results (%) of CoNTACT on each argument class per experiment.

	tw-tw			tw-fb			fb-fb			fb-tw			both-tw			both-fb		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
alt.	67	48	56	42	46	44	67	4	7	50	3	5	64	56	60	55	46	51
con.	60	49	54	32	23	27	55	31	40	52	31	39	57	47	51	54	43	48
crit.	51	34	41	49	42	45	61	57	59	25	46	33	49	37	42	59	58	59
dev.	58	47	52	56	24	36	55	33	41	56	36	44	58	51	54	54	34	42
eff	64	62	63	61	57	59	61	50	55	69	50	58	65	65	65	59	53	56
inst.	68	63	66	61	22	32	57	38	46	78	10	18	68	63	66	62	42	50
lib.	78	77	78	67	38	49	66	50	57	72	81	76	78	78	78	65	51	57
saf.	72	69	71	87	37	52	70	67	69	58	57	58	72	71	72	70	70	70
micro	70	64	67	62	36	46	64	53	58	58	51	55	69	65	67	63	56	59
macro	65	56	60	57	36	43	61	41	47	58	39	41	64	58	61	60	50	54
weighted	69	64	66	66	36	46	63	53	57	63	51	51	69	65	67	63	56	59
samples	58	57	56	37	34	34	55	51	51	53	48	48	59	58	57	56	54	53

In order to gain insight into the specific improvements of CoNTACT, a manual error analysis of the predictions of both models was conducted. First, instances where CoNTACT succeeded and RobBERT failed to predict the correct argument(s) were investigated. For each argument class, several terms seemed to guide the predictions of CoNTACT, because of the learned representations of said terms during both the domain adaptation and fine-tuning phase. For instance, references to the immune system and drugs, such as Ivermectine, were found to be stronger indicators of the "alternative medicine" argument class for CoNTACT than for RobBERT in predicting this argument. Further, comments containing words and hashtags such as "medical experiment" and "lab rat" were classified correctly by CoNTACT as related to "development", contrary to RobBERT, which made more false negative errors in this class. Similar observations were made for "institutional motives" (e.g. references to governments, political parties and politicians, such as #rutte3, #dv66 and #hugodejonge), "conspiracy theory" (e.g. references to gene therapy, such as "#geneticmodification"), "safety" (also references to gene therapy), and "liberty" (e.g. references to vaccine passports and obligation).

In addition, messages where RobBERT predicted the correct arguments but CoNTACT did not were investigated, although no clear error patterns were found in these cases. In general, however, both models seem to incorrectly classify arguments when the message itself lacks context or terminology related to the argument. For example, in the Facebook comment "they don't want them [the vaccines] anywhere else", which was annotated with the "criticism on vaccination strategy" label, both models failed to predict any argument, since the reference to e.g., a potential surplus of vaccines is implicit in this case.

In conclusion, CoNTACT seems to have learned domain-specific terminology in the domain adaptation phase, which benefits the model for the argument detection task, as can be derived from the results. The error analysis, however, showed that the model still experiences difficulties with classifying text entries that lack context or explicit information about the relevant argument(s).

Table 8. Statistically significant improvements in the argumentation detection task of CoNTACT over RobBERT.

Experiment	Classes with significant improvements
Tw - Tw	efficacy (***)
Tw - Fb	conspiracy (*), criticism on vaccination strategy (***), institutional motives (***), liberty (***)
Fb - Fb	development (***), efficacy (***), institutional motives (***), liberty (***), safety (*)
Fb - Tw	alternative medicine (***), conspiracy (***), criticism on vaccination strategy (***), development (***), efficacy (***), institutional motives (***), liberty (***), safety (*)
Both - Tw	efficacy (**)
Both - Fb	criticism on vaccination strategy (***), development (*)

5   Conclusion

In this work we presented CoNTACT, a Dutch language model adapted to the domain of COVID-19 tweets. The model was developed by continuing the masked language modeling pre-training phase of RobBERT using 2.8M Dutch tweets related to COVID-19. In order to test the performance of CoNTACT, the model was tested on two classification tasks: detection of vaccine hesitancy and detection of arguments for vaccine hesitancy. These tasked were performed in various experimental settings, that is by fine-tuning and testing on social media messages from two different platforms: Twitter and Facebook. For the vaccine hesitancy detection task, CoNTACT outperformed RobBERT with statistical significance in all experiments, including cross-genre settings. With respect to the argument classification task, CoNTACT showed substantial gains in virtually all classes in all experiments, some of which with statistical significance. An error analysis showed that the domain adaptation resulted in better representations of COVID-19 related terminology, and therefore in better results. Issues remain in messages containing implicit/figurative language or messages lacking context. Future work may include the development of a second version of CoNTACT, where the model is fine-tuned on more data from various platforms (Twitter, Facebook, Reddit, etc.) for even more cross-genre robustness.

References

1. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F.M., Rosso, P., Sanguinetti, M.: SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 54–63. Association for Computational Linguistics, Minneapolis (MN), USA (2019). https://doi.org/10.18653/v1/S19-2007

2. De Smedt, T., Daelemans, W.: Pattern for Python. J. Mach. Learn. Res. **13**(null), 2063–2067 (jun 2012)

3. Delobelle, P., Winters, T., Berendt, B.: RobBERT: a Dutch RoBERTa-based Language Model. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 3255–3265. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.292

4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR (2018), `http://arxiv.org/abs/1810.04805`

5. Durrani, N., Sajjad, H., Dalvi, F.: How transfer learning impacts linguistic knowledge in deep NLP models? CoRR (2021), `https://arxiv.org/abs/2105.15179`

6. Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H.R.: A brief review of domain adaptation. CoRR (2020), `https://arxiv.org/abs/2010.03978`

7. Han, X., Eisenstein, J.: Unsupervised domain adaptation of contextualized embeddings: A case study in early modern english. CoRR (2019), `http://arxiv.org/abs/1904.02817`

8. Joshi, A., Dai, X., Karimi, S., Sparks, R., Paris, C., MacIntyre, C.R.: Shot or not: Comparison of NLP approaches for vaccination behaviour detection. In: Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task. pp.

43–47. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/W18-5911

9. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: Compressing text classification models. CoRR (2016), `http://arxiv.org/abs/1612.03651`

10. Kreutz, T., Daelemans, W.: How to optimize your Twitter collection. Computational Linguistics in the Netherlands Journal **9**, 55–66 (Dec 2019), `https://www.clips.uantwerpen.be/clinjdraft/clinj/article/view/92`

11. Lemmens, J., Dejaeghere, T., Kreutz, T., Van Nooten, J., Markov, I., Daelemans, W.: Vaccinpraat: Monitoring vaccine skepticism in Dutch Twitter and Facebook comments. Computational Linguistics in the Netherlands Journal **11**, 173–188 (2021)

12. Leong, C.W.B., Beigman Klebanov, B., Hamill, C., Stemle, E., Ubale, R., Chen, X.: A report on the 2020 VUA and TOEFL metaphor detection shared task. In: Proceedings of the Second Workshop on Figurative Language Processing. pp. 18–29. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.figlang-1.3

13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. CoRR (2019), `http://arxiv.org/abs/1907.11692`

14. Martin, L., Müller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: CamemBERT: A tasty french language model. CoRR (2019), `http://arxiv.org/abs/1911.03894`

15. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947). https://doi.org/10.1007/bf02295996

16. Müller, M., Salathé, M., Kummervold, P.E.: COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. CoRR (2020), `https://arxiv.org/abs/2005.07503`

17. Rostami, M., Galstyan, A.: Domain adaptation for sentiment analysis using increased intraclass separation. CoRR (2021), `https://arxiv.org/abs/2107.01598`

18. Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E.: Deep learning for computer vision: A brief review. Computational Intelligence and Neuroscience pp. 1–13 (2018). https://doi.org/10.1155/2018/7068349

19. de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., Nissim, M.: Bertje: A Dutch BERT model. CoRR (2019), `http://arxiv.org/abs/1912.09582`

20. Wang, M., Deng, W.: Deep visual domain adaptation: A survey. CoRR (2018), `http://arxiv.org/abs/1802.03601`

21. Wang, S., Schraagen, M., Tjong Kim Sang, E., Dastani, M.: Public sentiment on governmental COVID-19 measures in Dutch social media. In: Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. Association for Computational Linguistics, Online (Dec 2020). https://doi.org/10.18653/v1/2020.nlpcovid19-2.17

22. Weinzierl, M., Harabagiu, S.: Automatic detection of covid-19 vaccine misinformation with graph link prediction. Journal of Biomedical Informatics **124**, 103955 (2021). https://doi.org/10.1016/j.jbi.2021.103955

23. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big data **3**(1), 9 (2016)

24. Xu, J., Xiao, L., López, A.M.: Self-supervised domain adaptation for computer vision tasks. CoRR (2019), `http://arxiv.org/abs/1907.10915`