Computational Linguistics in the Netherlands Journal 11 (2021) 173-188 Submitted 10/2021; Published 12/2021

Vaccinpraat: Monitoring Vaccine Skepticism in Dutch Twitter and Facebook Comments

Jens Lemmens* JENS.LEMMENS@UANTWERPEN.BE
Tess Dejaeghere* TESS.DEJAEGHERE@GMAIL.COM
Tim Kreutz* TIM.KREUTZ@UANTWERPEN.BE
Jens Van Nooten* JENS.VANNOOTEN@UANTWERPEN.BE
Ilia Markov* ILIA.MARKOV@UANTWERPEN.BE
Walter Daelemans* WALTER.DAELEMANS@UANTWERPEN.BE

* *University of Antwerp (CLiPS)*

Abstract

We present an online tool – "Vaccinpraat" – that monitors messages expressing skepticism towards COVID-19 vaccination on Dutch-language Twitter and Facebook. The tool provides live updates, statistics and qualitative insights into opinions about vaccines and arguments used to justify anti-vaccination opinions. An annotation task was set up to create training data for a model that determines the vaccine stance of a message and another model that detects arguments for anti-vaccination opinions. For the binary vaccine skepticism detection task (vaccine-skeptic vs. non-skeptic), our model obtained F1-scores of 0.77 and 0.69 for Twitter and Facebook, respectively. Experiments on argument detection showed that this multilabel task is more challenging than stance classification, with F1-scores ranging from 0.23 to 0.68 depending on the argument class, suggesting that more research in this area is needed. Additionally, we process the content of messages related to vaccines by applying named entity recognition, fine-grained emotion analysis, and author profiling techniques. Users of the tool can consult monthly reports in PDF format and request data with model predictions. The tool is available at `https://vaccinpraat.uantwerpen.be/`

1. Introduction

Expressions of mistrust and hesitancy towards vaccines, which have existed for centuries, increased substantially since the start of the COVID-19 crisis (Spinney 2021). Since social media platforms are the perfect instrument to spread opinions about vaccines (Smith et al. 2020), we present an online tool named "Vaccinpraat" (lit. "vaccine talk") to gain insights into topical COVID-19 vaccine opinions. Vaccinpraat monitors the stance towards COVID-19 vaccinations and the most frequently used anti-vaccination arguments on Dutch-language Twitter and Facebook by providing live statistics, content analyses and metadata analyses. An annotation task for vaccine stance and, in case of vaccine skepticism, the argumentation for this stance, was set up in order to obtain data for training supervised classification models that classify vaccine stance and (anti-vaccination) arguments in tweets and Facebook comments. Additionally, hashtags and named entities were extracted to investigate which subtopics are being discussed in messages expressing vaccine skepticism in more detail. Moreover, an emoji and emotion word analysis was conducted to provide an overview of the emotions that are expressed in vaccine-skeptic messages. Finally, Twitter users that post vaccine-skeptic messages were detected and geographically mapped to gain insights into which areas in Flanders and the Netherlands contain the most people with negative vaccine opinions. Authorship attribute predictions are also assigned to these users to form a more complete image of vaccine-skeptic Twitter users.

2. Related research

Joshi et al. (2018) provided an overview of existing approaches for vaccine behaviour classification, which is the task of determining whether a person has received or will receive a vaccine or not based on social media messages. Three main types of approaches were explored: rule-based, statistical, and deep learning approaches. Various models were designed for each type of approach and the performance of these models was compared. The results showed that both the pre-trained language model (test set F1-score of 80.43%) and the statistical ensemble model (test set F1-score of 81.56%) outperformed all other tested models (DNN, CNN, BiLSTM, rule-based, logistic regression with sentence vectors).

Skeppstedt et al. (2017) developed a classifier for detecting vaccine opinion (for, against or undecided) in discussion fora. A set of 1,190 vaccination related posts retrieved from Mumsnet[1], a British parental website where users can discuss parenting and related topics, was scraped and annotated for stance, and used to train a linear Support Vector Machines (SVM) classifier that uses token n-grams in two ways: first, the model was trained to classify the stance of the post (multiclass). Secondly, the model was trained to detect opinions against vaccines (binary classification). The model achieved F1-macro scores of 0.44 and 0.62 for the multiclass and binary classification tasks, respectively.

Other vaccination-related NLP research focused more on qualitative insights into vaccination opinions. Morante et al. (2020), for instance, created the Vaccination Corpus[2], which consists of 294 long-form text documents related to the 2015 measles outbreak in Disneyland, California that were retrieved from various online sources, such as blogs, government reports, (pseudo)science articles and news articles. All relevant vaccine-related events in the corpus were tagged, resulting in 6,722 unique events in total. Additionally, attribution relations were annotated. These attribution relations are a linguistic phenomenon where an attitude towards a target is ascribed to a third party, e.g. "The Hepatitis B vaccine [target] is considered one of the safest [attitude] vaccines by the WHO [third party]." Then, all opinions on (groups of) people and institutions, such as social media, conspiracy theorists and pharmaceutic companies were annotated. Finally, all arguments in the corpus were annotated (bottom-up) by indicating all claims and expressions of opinion concerning vaccinations. This task was the most challenging of all annotation tasks, because a clear formal definition of an argumentative claim is lacking, resulting in lower inter annotator agreement (IAA) than all other tasks.

Further research includes the identification of the different subtopics and fine-grained (misinformation) narratives within COVID-19 anti-vaccination messages. For example, a recent First Draft News[3] report distinguished 6 main anti-vaccination subtopics (Smith et al. 2020):

[1] **Development**: messages that express worry about the development, testing methodology, distribution, provision and public access of vaccines.
[2] **Safety, efficacy and necessity**: messages that state that vaccines are not safe, efficient (enough), or necessary to overcome the virus.
[3] **Institutional motives**: messages expressing mistrust in motives of political or economic institutions/people involved with vaccines.
[4] **Conspiracy theories**: messages that spread either well-established or completely new conspiracy theories.
[5] **Liberty and freedom**: messages that express concerns about how vaccines and the act of enforcing them upon people may affect civil liberty and personal freedom.
[6] **Morality and religion**: messages that show how the vaccine collides with moral or religious beliefs.

1. https://www.mumsnet.com/
2. https://github.com/cltl/VaccinationCorpus
3. The report and more information about First Draft News can be found here

These subtopics can be divided further into more fine-grained narratives after a qualitative analysis. For example, in the category "conspiracy theory", researchers found frequently recurring narratives such as "vaccines contain 5G", "Bill Gates is using the vaccines to control us" or "vaccines are used as a tool for depopulation programs". These narratives can then be broken down even further into individual utterances, i.e. actual social media messages.

In the First Draft News report, it was investigated for various languages (English, Spanish and French) how many messages related to each of the aforementioned subtopics could be found across different social media platforms (Facebook, Twitter and Instagram). Their main finding was that the distribution of these subtopics varies from language to language, similar to the different fine-grained narratives per subtopic. For example, messages pertaining to morality and religion were found substantially more frequently in Spanish messages than in French or English messages. Furthermore, it was found that different platforms dominated in different languages. Twitter, for instance, was substantially more dominant in English language messages than in French or Spanish messages. Vaccinpraat will provide insights into these matters in the case of Dutch-language social media.

3. Methodology

3.1 Data

We automatically scrape all Dutch-language tweets on a daily basis with the Twitter API using the method described by Kreutz and Daelemans (2019). Regular expressions are then used to filter all vaccination related messages. To obtain Facebook data, we use the Facepager[4] app (Jünger and Keyling 2019) to extract all posts originating from public pro-vaccination, anti-vaccination, and neutral pages on a weekly basis and scrape all comments to those posts that are publicly available. The Facebook pages were collected manually using the Facebook search function and boolean queries in Google search, and include pages from both Flanders and the Netherlands.

3.2 Preliminary stance classification

In the earlier stages of the tool, heuristics were used to determine the vaccine stance ("pro", "anti" or "neutral") of a message. For Twitter, we used hand-picked lists of pro- and anti-vaccination hashtags (see Table 9 in Appendix). In order to determine the vaccine stance, the following rules were employed:

- Tweets that contain neither pro- nor anti-vaccination hashtags are labeled as "neutral".
- Tweets that contain one or more anti-vaccination hashtag(s) are labeled as "anti-vaccination".
- Tweets containing one or more pro-vaccination hashtag(s) are labeled as "pro-vaccination".
- Since a manual analysis of the previous rule showed that false positives of the pro-vaccination class tend to occur in tweets where pro-vaccination hashtags are negated, tweets containing a pro-vaccination hashtag followed directly by "niet" ("not"), e.g. "#ikvaccineerniet" or "#ik-vaccineer niet", are labeled as "anti-vaccination".
- Further, we observed that tweets containing both pro- and anti-vaccination hashtags are often posted by vaccine skeptics criticizing people with pro-vaccination opinions. Therefore, these tweets were labeled as "anti-vaccination".

In the case of Facebook, we used the pages that comments were placed on as a proxy for their stance; comments on anti-vaccination pages received an "anti" label and comments on pro-vaccination pages were labeled as "pro". Comments originating from other pages were treated as neutral comments. Although this naive approach provides an indication of the "buzz" of COVID-19 vaccination on neutral, pro-vaccination, and anti-vaccination Facebook pages, an obvious drawback is that the

4. `https://github.com/strohne/Facepager`

Method	Class	Pre	Rec	F1
"Anti" baseline	Anti	0.33	1	0.50
	Pro	0	0	0
	Neutral	0	0	0
	Macro-averaged	0.17	0.17	0.17
Random baseline	Anti	0.33	0.33	0.33
	Pro	0.33	0.33	0.33
	Neutral	0.33	0.33	0.33
	Macro-averaged	0.33	0.33	0.33
Twitter heuristics	Anti	0.68	0.45	0.54
	Pro	0.49	0.76	0.60
	Neutral	0.45	0.40	0.45
	Macro-averaged	0.56	0.54	0.53
Facebook heuristics	Anti	0.66	0.55	0.60
	Pro	0.52	0.40	0.45
	Neutral	0.40	0.57	0.47
	Macro-averaged	0.53	0.51	0.51

Table 1: Performance of the heuristics and baselines on the stance classification task (measured on 300 manually annotated tweets and Facebook comments with balanced class distributions).

number of neutral comments is overestimated, because these pages (typically newspaper pages) are much more active than specific anti- or pro-vaccination pages, but by definition do not only contain neutral comments. To measure this overestimation of neutral comments and test the overall performance of the heuristics used for the preliminary stance classification, the methods were applied to 300 randomly selected annotated tweets and the same number of annotated Facebook comments with balanced class distributions (see Section 3.3 for a full description of the annotation process). Two baselines were used: one that predicted the "anti" label for all messages and another that randomly predicted one of the three classes.

The results, which can be found in Table 1, show that the heuristics for Twitter and Facebook perform (approximately) equally well and substantially better than the naive baselines for all of the classes, suggesting that they can be used for preliminary stance classification. As expected, however, an overestimation of the neutral class can be observed as evidenced by the low precision scores of that class in both Facebook and Twitter. Since it is clear that these results are sub-optimal and that more sophisticated models are necessary for high-quality predictions, an annotation task was set up to create training data for supervised models.

3.3 Annotation task

Three annotators (all Bachelor students and native speakers of Dutch) annotated the region, vaccine stance and argumentation (in the case of vaccine skepticism) of a subset of 8,855 vaccine-related tweets and 5,234 vaccine-related Facebook comments posted between December 2020 and May 2021. To avoid oversampling of neutral messages, the annotators were provided with data that contained 50% anti-vaccination messages, 25% pro-vaccination messages and 25% neutral messages as estimated by the heuristic methods. For Facebook, the posts of the comments were made available and if the comment was a reply to another comment, this comment was also provided as conversational context to aid the annotation process.

As evidenced by the aforementioned First Draft report, negative opinions towards vaccines can vary on a spectrum from slightly hesitant and questioning to radical paranoia and misinformation (Smith et al. 2020). Therefore, we distinguished between "vaccine hesitancy" and "anti-vaccination" messages during manual annotation, splitting the "anti" category used in our rule-based classification method into two more nuanced categories. For the region, possible labels were "Belgium (Flanders)"

Vaccine stance	Twitter	Facebook	Total
Anti-vaccination	4,424	2,194	6,618
Vaccine hesitancy	2,507	1,759	4,266
Neutral	579	396	975
Pro-vaccination	1,270	885	2,155
All	**8,780**	**5,234**	**14,014**

Table 2: Number of annotated messages per platform and vaccine stance.

Argument type	Twitter	Facebook	Total
Liberty	3,314 (1)	399 (4)	3,713
Institutional motives	1,704 (2)	276 (6)	1,980
Criticism on vaccination strategy	736 (4)	971 (2)	1,707
Safety	1,382 (3)	1,316 (1)	1,698
Efficacy	637 (5)	285 (5)	922
Development	450 (7)	467 (3)	917
Conspiracy theory	481 (6)	167 (7)	648
Morality	193 (8)	44 (8)	237
Alternative medicine	140 (9)	35 (9)	175

Table 3: Number of annotated anti-vaccination argument types per platform (descending order).

and "the Netherlands". Regarding the anti-vaccination arguments, we largely adopted the list of topics proposed in the First Draft Report described in Section 2, but we adapted this list to our specific needs; since a qualitative analysis of the collected data showed that messages frequently show criticism towards the government's vaccination policy, on the one hand, and talk about alternative medicine as a cure for COVID-19, on the other hand, we included these two categories as additional argument types. Further, we split the class "safety, efficacy and necessity" into "safety", on the one hand, and "efficacy and necessity", on the other hand, to better reflect the peculiarities of these two different categories. This resulted in a multiclass, multilabel classification task with 9 possible types of arguments (see Table 3 for a full overview of the different argument classes).

The annotators were provided with detailed annotation guidelines[5] containing definitions, examples and decision-making cues for both the stance and argument annotation tasks to ensure high quality gold standard labels. In order to train the annotators for the annotation tasks and calculate the inter-annotator agreement (IAA), a set of 400 Facebook comments and 400 tweets with a 50%/25%/25% distribution of anti-vaccination/pro-vaccination/neutral messages was sampled with the heuristics methods. For vaccination stance on Twitter and Facebook, respectively, the average pairwise Cohen's Kappa amounts to 0.3601 and 0.4967, whereas the average pairwise percent agreement is 52.00% and 64.85%. This relatively low IAA agreement was mostly caused by disagreement in the vaccine hesitancy and anti-vaccination classes: the average pairwise Cohen's Kappa for comments where all three annotators chose either "vaccine hesitancy" or "anti-vaccination" amounts to only 0.259. Because of this low agreement, the labels were converted to binary labels during the development of the supervised classification method described in Section 3.4 to avoid performance plateaus. Concretely, the labels "vaccine hesitancy" and "anti-vaccination" were regarded as "vaccine-skeptic" messages, whereas the group of neutral and pro-vaccination messages were considered as "non-skeptic". For these binary labels, the average pairwise Cohen's Kappa amounts to 0.5542 and 0.6630 (for Twitter and Facebook), and the average pairwise percent agreement to 82.30% and 82.76% (i.e. "moderate" to "substantial" agreement cf. Landis and Koch (1977)).

Since only the vaccine hesitancy and anti-vaccination messages are relevant for the argument labeling task, and these numbers are too low for separate evaluation, the IAA for this task was

5. https://github.com/clips/vaccinpraat_annotation_guidelines

calculated on both platforms together. Because argument annotation is a multilabel task, the annotators were asked to put the most dominant argument of a message as the first label and the IAA was based on these labels only. This resulted in an average pairwise Cohen's Kappa of 0.4369, whereas the average pairwise percent agreement amounted to 53.68%, corresponding to a "moderate" agreement (Landis and Koch 1977). After the training phase and the calculation of the IAA, it was decided to let the students annotate messages individually (one student per message) to maximize the number of annotations.

In Table 2, the number of annotated messages in Twitter and Facebook per stance can be found and Table 3 shows the number of annotated messages per anti-vaccination argument. From the statistics in these tables, it can be observed that the most frequently occurring vaccine stance in our subset is anti-vaccination, followed by vaccine hesitancy, pro-vaccination, and finally neutral comments. Note that the heuristics described in Section 3.2 were used to balance the data used for annotation and that the observed stance distribution therefore does not necessarily reflect the distribution of the data that is collected on a daily basis.

Regarding the arguments, the most to least frequent classes in our data are "liberty", "institutional motives", "criticism on vaccination strategy", "safety", "efficacy", "development", "conspiracy theory", "morality", and "alternative medicine" when considering both social media platforms simultaneously[6]. However, the order of this frequency varies between Twitter and Facebook: for Twitter, the most frequent arguments are "liberty", "institutional motives" and "safety", whereas "safety", "criticism on vaccination strategy" and "development" are observed the most frequently on Facebook. The argument classes "morality" and "alternative medicine" have the lowest frequencies in both Twitter and Facebook. Note that there is a large discrepancy in the frequency of certain arguments across the platforms. For example, "liberty" and "institutional motives" are arguments that occur substantially more frequently on Twitter than on Facebook in both absolute and relative terms. This could possibly be caused by the fact that we have access to all Dutch-language tweets posted daily, but not to all Dutch Facebook comments, since we are forced to select Facebook data manually and do not have access to private Facebook groups. Therefore, the presented (relative) frequencies of the arguments in the Facebook data may not represent the true distribution of arguments used in all Facebook comments.

3.4 BERT-based classification

The annotated data described in the previous section was used to develop a vaccine skepticism detection model and argument classification model based on BERTje/RobBERT (de Vries et al. 2019, Delobelle et al. 2020), the Dutch versions of BERT/RoBERTa (Devlin et al. 2019, Liu et al. 2019), respectively. During fine-tuning, it was investigated which parameter settings yielded highest results and whether fine-tuning on both social media platforms simultaneously improved results compared to fine-tuning only on messages scraped from a single platform. Further, it was examined whether adding conversational context to Facebook data can boost performance. This conversational context was added to the Facebook data by appending the previous turn in a comment thread (if any) and/or the post on which the comment was made to the end of relevant comment and separating them using the separator token ("[SEP]"). The conversational context was always put after the comment (most recent context first) so that when the entire string was longer than the maximum sequence length of the models, the context was truncated and not the message itself. For both classification tasks, a train-test split was constructed, splitting across post boundaries[7] for Facebook to prevent messages from the same discussion thread to appear both in the training and test sets, that is, to avoid within-post bias. The statistics of these splits can be found in Table 4 and 5.

6. Note that our data was collected in a specific time frame (December 2020 - May 2021) and that this distribution can vary over time.
7. Splitting across post boundaries makes it impossible to use k-fold cross-validation, i.e to create k splits of equal length with stratified label distributions, which is why a train-test split was used.

	Training			Testing		
	Twitter	**Facebook**	**Both**	**Twitter**	**Facebook**	**Both**
Vaccine-skeptic	6,238	3,596	9,834	693	321	1,014
Non-skeptic	1,664	1,144	2,808	185	117	302
Total	7,902	4,740	12,642	878	438	1,316

Table 4: The number of messages in the training and test sets for the stance classification task.

	Training			Testing		
	Twitter	**Facebook**	**Both**	**Twitter**	**Facebook**	**Both**
Alternative medicine	140	45	185	35	11	46
Conspiracy theory	550	176	726	137	52	189
Criticism vaccination strategy	783	910	1,693	196	312	508
Development	452	369	821	113	142	255
Efficacy	688	339	1,027	172	61	233
Institutional motives	1,455	212	1,667	364	100	464
Liberty	2,712	340	3,052	678	110	788
Morality	181	51	232	45	8	53
Safety	1,194	975	2,169	299	441	740
N messages	6,772	2,906	9,678	1,667	1,011	2,678

Table 5: The number of messages in the training and test sets for the argument classification task.

4. Results

4.1 Vaccine stance

For the classification of vaccine stance, RobBERT was used (Delobelle et al. 2020). It was investigated whether fine-tuning on Twitter and Facebook combined improved performance over fine-tuning on comments from these platforms separately. Further, the optimal number of epochs (3 to 10), batch size (4, 8, 16, 32, 64) and learning rate (3e-4, 1e-4, 5e-5, 3e-5, 2e-5) were grid searched, and it was determined whether using conversational context in Facebook boosts the performance (using the previous comment in the discussion thread, the post on which the comment was made, or both). The results were compared to a random baseline and to the heuristics methods described in Section 3.2. The Twitter heuristics, as shown in Table 6, performed considerably better in this binary classification setting (70% F1) than in the multiclass setting (53% F1), due to less confusion between the pro-vaccination and neutral classes. The Facebook heuristics, however, showed lower performance (30% F1), because only a small fraction of hesitant comments in the test set originated from anti-vaccination Facebook pages, an issue addressed earlier.

When examining the results in Table 6 further, it can be observed that the highest F1-scores obtained for Twitter and Facebook, respectively, are 0.77 and 0.69. Moreover, using conversational context and/or fine-tuning on both social media platforms simultaneously did not improve model performance on the stance classification task. Regarding the optimal hyperparameters, using 4 epochs for fine-tuning, a batch size of 8 and a learning rate of 5e-5 provided the best test set results for both platforms. Note that in the case where RobBERT was fine-tuned on tweets only, the optimal batch size increased to 64, although the batch size parameter had only marginal influence on the obtained results in this setting, meaning that the results with a batch size of 64 were not substantially better than with a batch size of 8. In addition, it was investigated whether RobBERT showed statistically significant improvements in performance compared to the heuristics, which was the case for both Twitter and Facebook, regardless whether RobBERT was trained on the data from a single or from both platforms ($p<0.001$ for all experiments; McNemar (1947)). For its implementation in Vaccinpraat, RobBERT was fine-tuned on all available annotated data (without conversational context) for 4 epochs using a batch size of 8 and a learning rate of 5e-5.

Train	Test	Epochs	Batch size	Learning rate	Context	Pre	Rec	F1
TW	TW	4	64	5e-5	N/A	0.78	0.76	0.77*
FB	FB	4	8	5e-5	None	0.69	0.69	0.69*
Both	TW	4	8	5e-5	None	0.75	0.75	0.75*
Both	FB	4	8	5e-5	None	0.69	0.69	0.69*
Both	Both	4	8	5e-5	None	0.73	0.73	0.73*
Heuristics	TW	N/A	N/A	N/A	N/A	0.69	0.72	0.70
Heuristics	FB	N/A	N/A	N/A	N/A	0.50	0.50	0.30
Random	TW	N/A	N/A	N/A	N/A	0.51	0.51	0.51
Random	FB	N/A	N/A	N/A	N/A	0.51	0.51	0.51

Table 6: Results (macro-averaged) for stance prediction, including the optimal parameter settings. The asterisk symbol indicates statistically significant gains over the heuristics methods.

4.2 Anti-vaccination arguments

For the detection of the anti-vaccination arguments, BERTje (de Vries et al. 2019) was used, since it yielded better results than RobBERT in preliminary experiments. Due to the low frequency of certain argument classes, we fine-tuned BERTje on both Facebook and Twitter messages simultaneously in order to benefit these classes. The "morality class", however, was still too underrepresented for the model to learn in preliminary experiments, which is why it was discarded in further experiments. The model's hyperparameters were optimized (exploring the same hyperparameters and hyperparameter settings as in the stance classification experiments described above), and it was investigated whether conversational context on Facebook improves the performance of the model. We report the performance of the model on the test set (and on the Facebook and Twitter comments in this test set separately).

In Table 7, the effect of conversational context on classification performance was investigated when fine-tuning on both Twitter and Facebook data. Macro-averaged F1-scores obtained on the entire test set, and on the tweets and Facebook comments in this test set separately are reported. As shown in Table 7, the best results on the Facebook data were obtained when using both the Facebook post and the previous conversational turn in the comment threads (if applicable) as conversational context. These results showed statistically significant improvements for three of the most frequent argument classes in Facebook: "safety", "development" and "criticism on vaccination strategy" ($p<0.001$ for all three classes, McNemar (1947)). For the tweets in the test set, the F1-score macro-averaged across classes could not be improved, although significant improvements in the "consipracy theory" ($p=0.01$), "institutional motives" ($p=0.02$), and "liberty" classes ($p<0.001$) were observed when using all conversational context items for Facebook comments, indicating cross-genre improvements in these frequent classes, but performance drops in the other (less frequent) classes (McNemar 1947). When considering the entire test set, best results were also obtained when using all conversational context, showing statistically significant gains for all classes except for Development ($p=0.03$ for "alternative medicine"; $p=0.007$ for "conspiracy theories" and "efficacy"; $p<0.001$ for "criticism on vaccination strategy", "liberty", "safety" and "institutional motives", McNemar (1947)). The optimal parameter settings for this experimental setting were 10 epochs for fine-tuning, a batch size of 16 and a learning rate of 2e-5.

In Table 8, the results of the individual argument classes can be found. We provide results on Twitter and Facebook separately, and on the entire test set, using the optimal amount of conversational context for that particular platform (determined in 7): when testing on tweets only, no additional context was used during training, when testing on Facebook comments only or the entire test set, all conversational context was used during training. Three baselines were provided: (1) a majority class baseline, which consistently predicted the most frequently-occurring argument class in our annotated data ("liberty" for Twitter and "safety" for Facebook), (2) a random baseline, and

Added context	Twitter (F1)	Facebook (F1)	Combined (F1)
None	**0.47**	0.35	0.44
Previous message	0.45	0.40	0.43
Post	0.44	0.42	0.45
All	0.45	**0.43**	**0.45**

Table 7: Test set results (macro-averaged) for the argument classification task with BERTje fine-tuned on both social media platforms, using various amounts of conversational context.

Argument	Twitter (F1)	Facebook (F1)	Both (F1)
Alternative medicine	0.46	0.35	0.43
Conspiracy theory	0.39	0.36	0.40
Criticism on vaccination strategy	0.30	0.55	0.41
Development	0.51	0.30	0.40
Efficacy	0.54	0.44	0.53
Institutional motives	0.23	0.34	0.23
Liberty	0.68	0.54	0.61
Safety	0.63	0.59	0.59
Macro-average F1	0.47	0.43	0.45
Micro-average F1	0.50	0.50	0.48
Hamming Loss	0.14	0.14	0.15
Majority baseline (macro-averaged)	0.07	0.08	0.05
Random baseline (macro-averaged)	0.15	0.13	0.14
SVM baseline (macro-averaged)	0.45	0.38	0.43

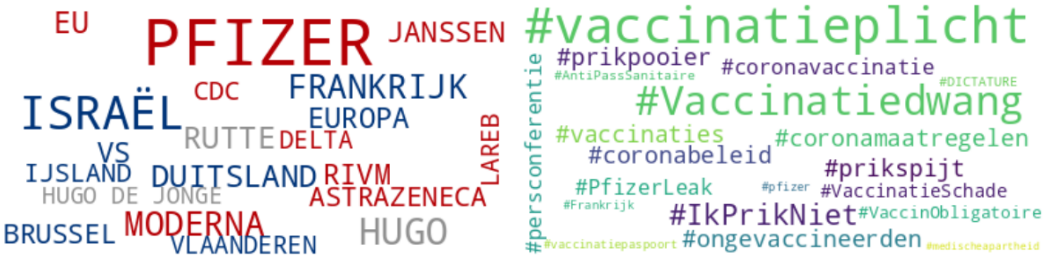
Table 8: Test results for argument detection with BERTje fine-tuned on both social media platforms (without the "morality" class; with added context items for Facebook comments).

(3) a binary relevance baseline that consisted of one SVM with linear kernel per class. Each of these SVM classifiers was trained to predict whether a specific argument was present in a message or not.

As shown by the results in Table 8, "institutional motives" were the most difficult to learn (0.23 F1), in spite of its high frequency in the training data. In contrast, the best classification performances were achieved for the "safety" (0.59 F1) and "liberty" (0.61 F1) classes due to their high frequency in the training data. Overall, macro-averaged F1 scores of 0.47 and 0.43 were obtained for Twitter and Facebook, respectively. For it's implementation in Vaccinpraat, BERTje was fine-tuned on all data (adding all conversational context to Facebook comments) for 10 epochs with a batch size of 16 and a learning rate of 2e-5, and it is used to detect arguments in messages from both platforms. The model will display the general trends of the most frequently occurring arguments correctly. However, the results for most classes are still sub-optimal, suggesting that more data is needed (especially for the underrepresented classes) in order to make the model more robust and represent trends in anti-vaccination arguments more accurately.

4.3 Qualitative and metadata analyses

In addition to the quantitative experiments described above, qualitative and metadata analyses were conducted to gain more insights into the subtopics that are being discussed in vaccine-skeptic messages, the emotions that are expressed the most frequently in these messages, and the Twitter users who posts these messages.



(a) Entities (red=org, blue=loc, grey=per) (b) Hashtags

Figure 1: Discussion topics in vaccine-skeptic messages.

4.3.1 Named entities and hashtags

The Dutch Stanza named entity recognizer (Qi et al. 2020) is used on Vaccinpraat to extract all people, locations, and organizations mentioned the most frequently in the vaccine-skeptic messages in the last 30 days. These named entities are then displayed in a word cloud, such as in Figure 1a, in which their size is weighted by their frequency. It can be observed that the most frequently mentioned organizations at the time this word cloud was produced (26 August 2021) were pharmacies such as Pfizer, Moderna, Astrazeneca, and Johnson and Johnson ("Janssen"). The most frequently discussed locations, on the other hand, were France ("Frankrijk"), due to the severe corona measurements that were taken in France at this time, and Israel, because of how quickly the adult population got vaccinated there. The most frequently mentioned people were Mark Rutte (the Dutch Prime Minister) and Hugo De Jonge (the Dutch Minister of Health), who are often criticized in vaccine-skeptic messages for their policies.

In addition, all hashtags in vaccine-skeptic tweets posted in the last 30 days are extracted and collected in a word cloud, which is displayed on Vaccinpraat to give a more comprehensive overview of the subtopics that are being discussed the most. As shown in Figure 1b (generated 26 August 2021), the most frequently used hashtags in this particular period were "#vaccinatieplicht" ("#vaccinationobligation") and "#vaccinatiedwang" ("#vaccinationcoercion"), indicating that personal freedom is an important discussion topic (which is also evidenced by the high frequency of the "liberty" argument class in our annotated data).

4.3.2 Emotions

Besides anti-vaccination arguments and topics, Vaccinpraat also provides insights into the emotions that are expressed in vaccine-skeptic messages. To detect the most dominant emotion in each message, the Dutch LiLaH[8] emotion lexicon was used (Ljubešić et al. 2020, Daelemans et al. 2020). This lexicon features the following 8 emotions: anger ("woede"), fear ("angst"), disgust ("afkeer"), joy ("plezier"), sadness ("droefheid"), trust ("vertrouwen"), anticipation ("anticipatie") and surprise ("verbazing"). For each of these emotions, all words in a message related to that particular emotion were counted. Then, the emotion with the highest word count was treated as the most dominant emotion for this message. As shown in Figure 2, this information is used to show the frequencies of the emotions expressed in vaccine-skeptic messages on a weekly basis. For weeks 31, 32 and 33 of 2021, for instance, the most dominant emotions were (mis)trust, anticipation and fear, whereas the least dominant emotions were surprise, disgust and joy. The latter emotion, joy, typically manifests in expressions such as "I am happy that I did not get vaccinated". Overall, we observe that the relative frequencies of the emotions remain stable on a week-to-week basis.

8. Linguistic Landscape of Hate Speech in Social Media: `https://lilah.eu/`

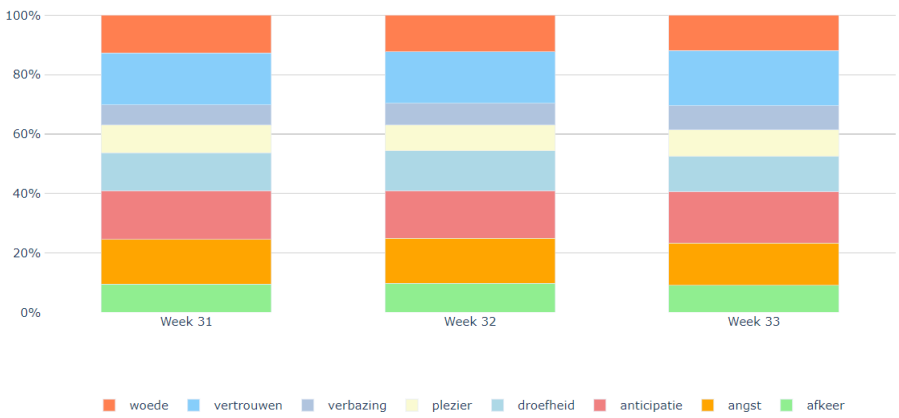


Figure 2: Relative frequencies of the most dominant emotion per message in weeks 31-33 of 2021.

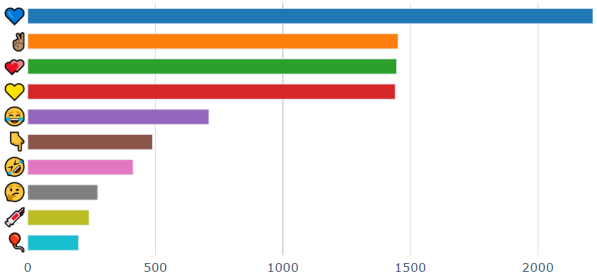


Figure 3: Absolute frequencies of the most frequently occurring emoji in vaccine-skeptic messages.

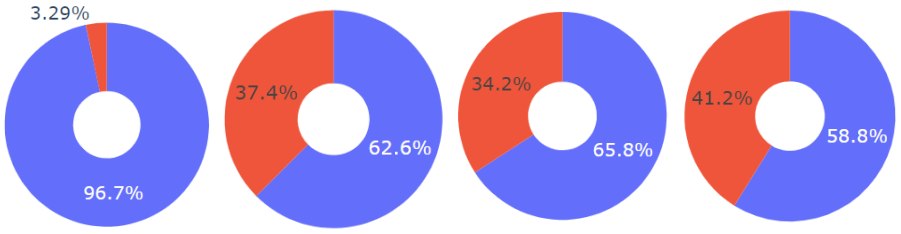
To complement this emotion word analysis, we collect the emoji occurring in vaccine-skeptic messages in Figure 3. As shown in this figure, varieties of heart emoji, smiling emoji and a syringe emoji can be found in the top 10 most frequently used emoji in vaccine-skeptic messages. A qualitative, manual analysis showed that positive emoji such as smiling/laughing emoji are often used sarcastically or ironically, similar to the clapping hands emoji, which are also used to approve anti-vaccination statements of others. The pointing finger emoji, on the other hand, are used to refer to hyperlinks, such as other tweets or news articles.

4.3.3 Metadata

To determine who posts vaccine-skeptic messages on social media, the geographical location of the Twitter users is determined by extracting the location of their recently posted tweets with the Twitter API. Then, these locations are mapped using the Google Maps API. Afterwards, author attributes, that is age, gender, personality and education level, are assigned to the users in our database by using the authorship attribute tools in the Textgain API (`https://www.textgain.com/product/api/`), which have obtained test set F1-scores of 75% (age), 85% (gender), 67% (personality) and 81% (education level). To increase the quality of their predictions, the models are run over the last 9 tweets of a user and majority voting is used to determine the final labels. All models base their predictions on text alone, with the exception of the gender detection model, which also uses the first name of the user to make predictions. Please refer to van de Loo et al. (2016) and Textgain's introduction page[9] for more information about the authorship attribution models.

9. `https://www.textgain.com/textgain-has-launched/`

Figure 4: Heatmap of Twitter users with anti-vaccination opinions.



(a) Age (blue=adult, red=adolescent

(b) Gender (blue=male, red=female)

(c) Personality (blue=extrovert, red=introvert)

(d) Education level (blue=low, red=high)

Figure 5: Authorship attributes of vaccine-skeptic Twitter users.

The geographical location of the vaccine-hesitant Twitter users can be found in the heatmap presented in Figure 4. Note that this figure is based on absolute numbers (not relative to the total number of Twitter users or the number of inhabitants). As can be observed from this figure, most vaccine skeptics in the Netherlands can be found around Amsterdam, Den Haag, Utrecht and Rotterdam, whereas for Flanders, most vaccine skeptics are located around Antwerp and Brussels, not surprisingly given the larger numbers of inhabitants in these cities compared to others. Nevertheless, these hotspots also show lower relative vaccine coverage in adults compared to the rest of their respective countries[10].

In Figure 5a, 5b, 5c and 5d, respectively, the distribution of the age (adult/adolescent), gender (male/female)[11], personality (introvert/extrovert) and education level (high/low) can be found. Note that these numbers are, similarly to the geographical data, not normalized by the total number of Twitter users. The aforementioned figures show that the vaccine skeptics on Twitter we found are often male (62.6%), mature (96.7%), extroverts (65.8%) and lower educated (58.8%), which is in line

10. `https://www.laatjevaccineren.be/vaccinatieteller-cijfers-per-gemeente`; `https://www.volksgezondheidenzorg.info/onderwerp/vaccinaties/regionaal-internationaal/covid-19`
11. The employed model is limited to binary labels only.

with previous research on the correlation between vaccine hesitancy and education level[12]. Since the metadata presented in this section is based on absolute numbers, i.e. not relative to the total number of Twitter users, it should be interpreted with caution before drawing direct conclusions. We acknowledge this limitation of the monitor and hope to address it in future research. Nevertheless, the metadata may indicate certain trends and shifts over time.

5. Conclusion

In this paper, we presented an online tool for monitoring vaccine opinions and anti-vaccination arguments in Dutch-language Twitter and Facebook. Vaccine-related messages were extracted using regular expressions and pre-classified using heuristics, which achieved reasonable performance when compared to a random baseline. Further, an annotation task was set up to label vaccine-related messages with stance and argumentation in case of vaccine skepticism. Qualitative analyses of vaccine-skeptic comments suggest that negative vaccine stances can range from nuanced and slightly hesitant to radical anti-vaccination opinions and belief in conspiracy theories. However, the annotation task and inter-annotator agreement show that it is non-trivial even for trained human annotators to distinguish between slightly hesitant and radical anti-vaccination opinions.

Our binary vaccine skepticism classifier yields performances of 0.77 and 0.69 F1-macro on Twitter and Facebook, respectively, and outperforms the heuristics methods significantly. However, attempts to improve performance by adding conversational context failed for this task. For argument detection, on the other hand, we show that adding conversational context can significantly improve classification performance for most classes. Nevertheless, the task itself of determining the arguments in vaccine-skeptic messages is complex, even for human annotators (as shown previously in (Morante et al. 2020)), suggesting that more research into methods that can extract arguments from opinions is needed. Another important challenge is that of unbalanced data, which could be tackled in future work by investigating which key words can be used to preselect and balance data more efficiently.

Finally, our vaccine monitor sheds light on which people, organisations, locations and hashtags are mentioned the most frequently in vaccine-skeptic messages, which in combination with the argument detection provides a comprehensive overview of the content that is discussed in today's anti-vaccination discourse. The results show that arguments related to politics and politicians, personal freedom, and the safety of the vaccines are dominant in current (anti-)vaccination discourse, whereas the most dominant emotions expressed in vaccine-skeptic messages are (mis)trust, anticipation and even fear.

In sum, Vaccinpraat provides multidimensional insights into current vaccine opinions, anti-vaccination arguments and emotions. These insights are crucial in times of a global pandemic, since they provide the first step towards understanding the concerns of vaccine skeptics and to unify people with positive and negative vaccine opinions. Future research directions in this field may include exploring new key word methods to collect more balanced data, and new annotation (and classification methods that quantify vaccine hesitancy on a scale rather than binarize it).

6. Acknowledgements

This research was partially funded by the Belgian Ministry of Health (FN459900017). We would like to thank the Department of Communication Sciences and the Center for the Evaluation of Vaccination with whom we cooperate in the Vaxcom research group. Additionally, we would like to thank Textgain for providing their authorship attribution tools.

12. https://www.persinfo.org/nl/nieuws/artikel/vub-onderzoek-toont-dat-meerderheid-bevolking-positief-staat-tegenover-vaccinatie/48800

References

Daelemans, Walter, Darja Fišer, Jasmin Franza, Denis Kranjčić, Jens Lemmens, Nikola Ljubešić, Ilia Markov, and Damjan Popič (2020), The LiLaH Emotion Lexicon of Croatian, Dutch and Slovene. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1318.

de Vries, Wietse, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim (2019), BERTje: A Dutch BERT model, *CoRR*. http://arxiv.org/abs/1912.09582.

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2020), RobBERT: A Dutch RoBERTa-based language model, *CoRR*. https://arxiv.org/abs/2001.06286, archivePrefix = arXiv.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://aclanthology.org/N19-1423.

Joshi, Aditya, Xiang Dai, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C Raina MacIntyre (2018), Shot or not: Comparison of NLP approaches for vaccination behaviour detection, *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, Association for Computational Linguistics, Brussels, Belgium, pp. 43–47. https://www.aclweb.org/anthology/W18-5911.

Jünger, Jakob and Till Keyling (2019), Facepager: An Application for Automated Data Retrieval on the Web.

Kreutz, Tim and Walter Daelemans (2019), How to optimize your Twitter collection, *Computational Linguistics in the Netherlands Journal (CLIN)* **9**, pp. 55–66. https://www.clinjournal.org/index.php/clinj/article/view/92.

Landis, J. Richard and Gary G. Koch (1977), The measurement of observer agreement for categorical data, *Biometrics* pp. 159–174, JSTOR.

Liu, Yinhan, 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. https://arxiv.org/abs/1907.11692.

Ljubešić, Nikola, Ilia Markov, Darja Fišer, and Walter Daelemans (2020), The LiLaH emotion lexicon of Croatian, Dutch and Slovene, *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, Association for Computational Linguistics, Barcelona, Spain (Online), pp. 153–157. https://aclanthology.org/2020.peoples-1.15.

McNemar, Quinn (1947), Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* **12** (2), pp. 153–157, Springer Science and Business Media LLC. https://doi.org/10.1007/bf02295996.

Morante, Roser, Chantal van Son, Isa Maks, and Piek Vossen (2020), Annotating perspectives on vaccination, *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 4964–4973. https://aclanthology.org/2020.lrec-1.611.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning (2020), Stanza: A Python natural language processing toolkit for many human languages, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

Skeppstedt, Maria, Andreas Kerren, and Manfred Stede (2017), Automatic detection of stance towards vaccination in online discussion forums, *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, Association for Computational Linguistics, Taipei, Taiwan, pp. 1–8. https://aclanthology.org/W17-5801.

Smith, Rory, Seb Cubbon, and Claire Wardle (2020), Under the Surface: COVID-19 Vaccine Narratives, Misinformation and Data Deficits on Social Media. First Draft, https://firstdraftnews.org/wp-content/uploads/2020/11/FirstDraft_Underthesurface_Fullreport_Final.pdf?x48126.

Spinney, Laura (2021), Could understanding the history of anti-vaccine sentiment help us to overcome it?, *The Guardian*. https://www.theguardian.com/society/2021/jan/26/could-understanding-the-history-of-anti-vaccine-sentiment-help-us-to-overcome-it.

van de Loo, Janneke, Guy De Pauw, and Walter Daelemans (2016), Text-based age and gender prediction for online safety monitoring, *International Journal of Cyber-Security and Digital Forensics (IJCSDF)* **5** (1), pp. 46–60.

Appendix

Anti-vaccination hashtags	Pro-vaccination hashtags
#vaccinatiedwang	#ikvaccineer
#vaccinatieplicht	#ikvaccineerwel
#ikvaccineerniet	#iklaatmevaccineren
#vaccinvrij	#ikprikwel
#vaccinatienazi	#ikwildieprik
#verplichtevaccinatie	#ikprik
#kritischprikken	#ikprikhetwel
#vaccinschade	#govaccyourself
#vaccinatieschade	#provaccinatie
#baasineigenlijf	#laatjevaccineren
#hugodejongekanniks	#vaccineer
#vaccinatiedwangnevernooitniet	#vaccineerdezorg
#overheidsdwang	#vaccineerdeouderen
#overheidspropaganda	#ikwilhetvaccin
#vaccinatie-apartheid	#ikwileenvaccin
#integriteitlichaam	#spuitmevol
#engeldesdoodsdejonge	#nlvaccineert
#geengewoonvaccin	#ikhebdieprik
#gezondheidsdicatuur	#ikvaccineerme
#vaccinatieapartheid	#vaccineernu
#vaccinatienazi's	#ikvaccineerwel
#vaccinatiebewijs	#ikdoewelmee
#vaccinatiepaspoort	
#vrijheidvooreenprikkie	
#testdwang	
#neetegenvaccinatie	
#nietvaccineren	
#vaccinatiedoden	
#ikprikhetniet	
#vaccinazi	
#vaccinatiepas	
#vaccinpas	
#vaccinpaspoort	
#vaccinbewijs	
#vaccindwang	
#vaccinplicht	
#geenvaccin	

Table 9: Anti- and pro-vaccination hashtags used in rule-based vaccine stance classification in tweets