
Proceedings Paper:
Li, Y., Scarton, C., Song, X. et al. (2023) Classifying COVID-19 vaccine narratives. In: Mitkov, R. and Angelova, G., (eds.) Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. 14th International Conference on Recent Advances in Natural Language Processing (RANLP 2023), 04-06 Sep 2023, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria, pp. 648-657. ISBN: 978-954-452-092-2. ISSN: 2603-2813.

https://doi.org/10.26615/978-954-452-092-2_070

Classifying COVID-19 Vaccine Narratives

Yue Li, Carolina Scarton, Xingyi Song and **Kalina Bontcheva**
Department of Computer Science, University of Sheffield (UK)
{yli381, c.scarton, x.song, k.bontcheva}@sheffield.ac.uk

Abstract

Vaccine hesitancy is widespread, despite the government's information campaigns and the efforts of the World Health Organisation (WHO). Categorising the topics within vaccine-related narratives is crucial to understand the concerns expressed in discussions and identify the specific issues that contribute to vaccine hesitancy.

This paper addresses the need for monitoring and analysing vaccine narratives online by introducing a novel vaccine narrative classification task, which categorises COVID-19 vaccine claims into one of seven categories. Following a data augmentation approach, we first construct a novel dataset for this new classification task, focusing on the minority classes. We also make use of fact-checker annotated data. The paper also presents a neural vaccine narrative classifier that achieves an accuracy of 84% under cross-validation. The classifier is publicly available for researchers and journalists.

1 Introduction

Vaccination is one of the most effective public health interventions, but it is essential that immunisation programs are able to achieve and sustain high vaccine uptake rates. Overcoming vaccine hesitancy, which refers to the delay in the uptake or refusal of vaccines, is a major challenge (Eskola et al., 2015) and the WHO has named it one of the top ten threats to global health in 2019 (Qayum, 2019). Vaccine hesitancy is a complex and context specific phenomenon, varying across time, place and even vaccines (Larson et al., 2014). It could be caused by various factors such as concerns about side effects, costs, and misinformation.

Although social media platforms like Twitter, Facebook, and YouTube have taken actions to limit the spread of misinformation, simply identifying and removing misinformation from platforms is not enough, as the concerns of the vaccine-hesitant citizens also need to be monitored and responded to. Consequently, fact-checkers and other professionals need analytical tools that help them to better monitor misinformation, vaccine hesitancy, vaccine-related debates and their narratives.

Topic analysis of narratives about vaccines could be used for this purpose, however, a large manual effort is required, due to the lack of a vaccine-related topic classifier. For example, Smith et al. (2020) gather over 14 million vaccine-related posts from Twitter, Instagram, and Facebook to research vaccine-related narratives. The posts are categorised into six topics based on a novel typology designed to capture the ways narratives are framed. However, manual analysis was feasible on only a small sample of 1,200 posts, which, given the small scales, leaves significant gaps in the understanding and tackling of vaccine hesitancy.

Guided by these needs, the novel contributions of this paper are in:

1. **Proposing a new seven-way classification task and dataset** for categorising vaccine related online narratives. The classification task adopts the six categories (see Table 1) defined by Smith et al. (2020). The dataset is built based on manual annotation and data augmentation [1]. Our experiment demonstrates that the augmented data significantly boosts classifier performance.

2. **Building and making available a vaccine narrative classifier**, based on the Classification Aware Neural Topic Model (CANTM)(Song et al., 2021). CANTM originally achieved state-of-the-art performance in COVID-19 misinformation classification

[1]We release the newly collected Twitter data: doi:10.5281/zenodo.8192131

Proceedings of Recent Advances in Natural Language Processing, pages 648–657
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_070

(Song et al., 2021) and is particularly suited to vaccine narrative classification too, as it is robust on small training sets. For reproducibility, the classifier is publicly available as a web service [2].

2 Related Work

Since the outbreak of the COVID-19 pandemic and accompanying infodemic, large-scale monolingual and multilingual datasets have been collected from different social media platforms in order to intervene and combat the spreading of COVID-19-related disinformation (Shuja et al., 2021; Alam et al., 2021; Shahi and Nandini, 2020; Li et al., 2020; Zarei et al., 2020), with vaccines being a commonly included topic in these datasets. As the importance of understanding and tackling COVID-19 vaccination hesitancy grew, increasing efforts have been made to analyse vaccine narratives and discourses, the dissemination of false claims and the anti-vaccine groups on social media. This has resulted in the construction of a number of COVID-19 vaccine-focused datasets, without (DeVerna et al., 2021; Muric et al., 2021) or with annotations about veracity (e.g., true or false information) (Hayawi et al., 2022), sentiment (e.g., positive, negative or neutral) (Kunneman et al., 2020), stance (e.g., pro- or anti-vaccine) (Mu et al., 2023; Agerri et al., 2021; Argyris et al., 2021) or topic category (e.g., vaccine development or side effects) (Bonnevie et al., 2021; Smith et al., 2020). The datasets, consequently, can be used to facilitate the research on COVID-19 vaccine-related online information from different aspects, including fact-checking, sentiment analysis, stance detection, and topic analysis.

Topics or themes discussed in the vaccine-related narratives and online debates are an essential dimension. State-of-the-art methods for automatic topic analysis typically fall under one of these categories: topic modelling (Jamison et al., 2020; Lyu et al., 2021; Chen et al., 2021; Xue et al., 2020), clustering (Sharma et al., 2022; DeVerna et al., 2021; Muric et al., 2021; Argyris et al., 2021), and inductive analysis (Bonnevie et al., 2021; Smith et al., 2020). Topic modelling, represented by Latent Dirichlet Allocation (LDA) (Blei et al., 2003), is the most commonly used approach at present (Jamison et al., 2020; Lyu et al., 2021; Chen et al., 2021; Xue et al., 2020). Clustering methods for topic discovery have been applied to text representations (Sharma et al., 2022; Smith et al., 2020) or networks (DeVerna et al., 2021; Muric et al., 2021). For instance, K-means (Lloyd, 1982) has been used to cluster the average word embeddings of vaccine narratives (Sharma et al., 2022) or to test a human-derived topic typology (Smith et al., 2020). After constructing a co-occurrence topic network with hashtags as nodes, the Louvain method (Blondel et al., 2008) is used to extract clustering from the graph (DeVerna et al., 2021; Muric et al., 2021). The above methods are unsupervised, resulting in no control on the model generation. Therefore, extra work is normally involved in discovering and labelling the topics.

In contrast, inductive analysis relies on experts to analyse the raw textual data and derive topics or themes (Bonnevie et al., 2021; Hughes et al., 2021; Smith et al., 2020). For instance, Bonnevie et al. (2021) categorise anti-vaccine tweets into twelve conversation themes, such as `negative health impacts`, `pharmaceutical industry` and `religion`. Hughes et al. (2021) identify twenty-two narrative tropes (e.g., `corrupt elites` and `vaccine injury`) and sixteen rhetorical strategies (e.g., `brave truthteller` and `appropriating feminism`) in anti-vaccine and COVID-denialist social media posts.

Besides the above work specific to anti-vaccine contents, general COVID-19 vaccine narratives on social media were categorised by fact-checkers and researchers at First Draft (Smith et al., 2020) as belonging to one of six topics, as shown in Table 1.

A potential drawback of inductive analyses is that the amount of data that can be analysed by the human experts is significantly smaller than the volumes analysed through the automatic topic modelling and clustering methods. To overcome this problem, Bonnevie et al. (2021) create a list of unique keywords for each theme during inductive analysis, which are then used to automatically categorise more posts based on keyword matching.

In this paper, we explore machine learning and deep learning methods for automatic vaccine narrative classification according to the topics proposed by Smith et al. (2020).

To the best of our knowledge, this is the first paper to frame online vaccine narrative categorisation as a classification task. In that respect,

[2] https://cloud.gate.ac.uk/shopfront/displayItem/covid19-vaccine

Topic	Description	Examples
Conspiracy (Cons)	Known or novel conspiracies and conspiracy theories involving vaccines or their development	Bill Gates: We need to depopulate the planet. Also Bill Gates: Save your life with my vaccine.
Development, Provision and Access (DPA)	The ongoing progress or challenges concerning the development, testing and provision of vaccines as well as the access to vaccines	Oxford coronavirus vaccine triggers immune response.
Liberty/Freedom (LF)	Civil liberties and personal freedom considerations surrounding vaccines and vaccination policies	States have authority to fine or jail people who refuse coronavirus vaccine, attorney says.
Morality, Religiosity and Ethics (MRE)	Moral, ethical and religious concerns around vaccines	Kanye West Praises Trump, Hammers Planned Parenthood, Likens COVID Vaccine To 'Mark Of The Beast'.
Politics and Economics (PE)	Political, economic or business developments related to vaccines	Scientists Worry About Political Influence Over Coronavirus Vaccine Project.
Safety, Efficacy and Necessity (SEN)	Safety and efficacy of vaccines, including the perceived necessity of vaccines	WHO warns coronavirus vaccine alone won't end pandemic: 'We cannot go back to the way things were'.

Table 1: Description and examples of each topic.

there are two closely relevant studies. Song et al. (2021) collect English debunks about COVID-19 and annotate them with ten disinformation categories. They also propose a novel framework that combines classification and topic modelling. Similarly, Shahi and Nandini (2020) scrape multilingual COVID-19 related fact-check articles and manually classify them into eleven topics, but the models they explore are limited to veracity prediction. Both papers study disinformation regarding COVID-19, with vaccine covered as only one monolithic category (`vaccines, medical treatments, and tests` (Song et al., 2021) or `prevention & treatments` (Shahi and Nandini, 2020)). However, our work is vaccine-focused, aiming at finer-grained, automatic categorisation of vaccine narratives.

3 Vaccine Narrative Categorisation: Task Definition and Dataset Construction

3.1 Definition

We define the COVID-19 vaccine narrative categorisation task as assigning COVID-19 vaccine-related claims to one of the six target topics identified by Smith et al. (2020): (1) `Cons` for vaccine-related conspiracies; (2) `DPA` for development, provision, and access to vaccination; (3) `LF` for vaccine-related civil liberties and freedom of choice; (4) `MRE` for moral, religious, and ethical concerns; (5) `PE` for political, economic, or business aspects; and (6) `SEN` for safety and efficacy concerns.

More detailed definitions and examples of the six topics are shown in Table 1.

In addition, we introduce a new, seventh category that encompasses claims related to animal vaccines (`AnimalVac`). The motivation is to recognise or filter out animal vaccine-related posts, which are also captured by keyword-based data collection methods that are typically used for collecting vaccine-related social media posts (e.g., using keywords such as vaccine or vaccines).

Thus, this paper regards the vaccine narrative categorisation task as a seven-way classification problem, with six topics pertaining to COVID-19 human vaccination and one additional topic for animal vaccination.

3.2 Dataset Construction

3.2.1 FD data

First Draft researchers and journalists (FD data) collected and manually annotated a number of posts in English with the six human vaccine related topics by Smith et al. (2020). Focusing on COVID-19 vaccine, the data covers general vaccine narratives, rather than only misinformation. It is gathered from multiple online platforms (news media, Twitter, Facebook, and Instagram), consisting of texts, images, and videos.

For our experiments all duplicates were removed, together with posts having just video content, since our aim is text-based classification. Posts with images are classified on the basis of their textual content if available and the alternative/alt texts [3] accompanying the images.

Table 2 shows the topic distribution of the English FD dataset after data filtering is applied.

3.2.2 Data Augmentation

As shown in Table 2, the FD dataset is highly imbalanced. `Cons`, `LF`, and `MRE` are minority classes, which only contain 6%, 9%, and 2% of the total posts, respectively. Besides, the FD dataset does

[3] a short written description of an image, which describes that image for accessibility reasons

	Cons	DPA	LF	MRE	PE	SEN	AnimalVac
FD data	26(6%)	116(27%)	37(9%)	7(2%)	108(25%)	134(31%)	0(0%)
Augmented	107(13%)	116(14%)	92(12%)	151(19%)	108(13%)	134(17%)	96 (12%)

Table 2: Distribution of data between classes before and after data augmentation.

not contact posts pertaining to animal vaccines, as these were excluded during their manual analysis.

To address these issues, we perform data augmentation, which includes the collection of new posts for the AnimalVac class, as well as gathering more examples for the three under-represented categories.

Using the Twitter API, we collected posts with vaccine-related hashtags such as #covidvaccine, #AstraZeneca, #vaccines. These tweets are then filtered on the basis of class-specific keywords and hashtags which we identified manually for each target class. As we aim to limit the overlap between the FD dataset and our newly collected data, we derived the keywords and hashtags on the basis of the FD codebook, i.e. annotator guidelines:

Cons: known conspiracy theories are considered, such as QAnon, ID2020, nanorobots insertion, new world order, and deep state. In addition, we included two other conspiracies fact-checked by the International Fact Checking Network (IFCN) [4], but not captured in the FD data: (a) The body can receive 5G signal after the vaccine is taken; and (b) China is collecting human DNA from all over the world through its vaccines in order to create a biological weapon.

LF: hashtags and terms regarding mandatory vaccination (e.g., #MandatoryVaccine, #NoJabNoPay), and concepts suggesting that mandatory vaccine programs undermine personal liberty or constitute a medical dictatorship (e.g., #MedicalFreedom, #InformedConsent, #MyBodyMyChoice).

MRE: keywords about how people are being used as animals in vaccine testing (e.g., lab rats, guinea pigs), and about religion or ideological stance in opposition to vaccines (e.g., aborted fetuses, changing DNA).

AnimalVac: hashtags such as #animalhealth, #WorldAnimalVaccinationDay, and #petmedicine are utilised to find the target tweets. As the number of the matched tweets is relatively small, we also collect Facebook posts to balance the dataset. They are picked out if they contain certain names of animal diseases and the word "vaccine".

The full list of keywords and hashtags per class are shown with examples in Table 3. All posts matching the keywords and hashtags for each target class are then manually annotated by the authors, in order to ensure label quality. Table 2 also presents the new data distribution following this augmentation. The proportion of Cons, LF and MRE has increased to 13%, 12%, and 19% respectively and 96 posts related to animal vaccines are also included.

4 Predictive Model

We evaluate feature-based and transformer-based models that are pre-trained with out-of-domain and in-domain data, and models that combine classification and topic modelling.

BOW-LR: We train a Logistic Regression model with bag-of-words using L2 regularisation, using the scikit-learn implementation (Pedregosa et al., 2011).

SCHOLAR: (Card et al., 2018) Sparse Contextual Hidden and Observed Language AutoencodeR adopts VAE and directly inserts label information in the encoder during training in order to generate latent variables dependent on the labels. Zero vectors are used to represent the labels in the test set during inference. We use the author's implementation of SCHOLAR (https://github.com/dallascard/scholar).

CANTM and CANTM-COVID (Song et al., 2021): Classification-Aware Neural Topic Model achieves state-of-the-art performance on COVID-19 disinformation categorisation (Song et al., 2021). It overcomes the shortage of SCHOLAR that the label information is unavailable during inference by designing a stack of two classifier-aware VAEs. The input text is first encoded by a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019), and a classifier is jointly trained with one of the VAEs, whose generated latent variables is the input of this classi-

[4] https://www.poynter.org/coronavirusfactsalliance/

Class	Keywords/hashtags	Examples
Cons	QAnon, new world order, nano, ID2020, deep state, China weapon, China DNA, 5g	(1) Vaccination day. When the time comes, get vaccinated. No one will microchip you like a cat and 5G will not control your mind. (2) Filled with nano particles to alter our DNA! The Moderna vaccine is the Gates vaccine.
LF	#freedom, #liberty, #NoVaccineForMe, #MyBodyMyChoice, #InformedConsent, #MandatoryVaccine, #MedicalFreedom, #NoJabNoPay, medical dictatorship, mandatory	(1) Before you all start, this is NOT about Pro #Vaccination or those against. This is about how the #nojabnopay discriminates against free choice and the rich/poor. (2) This is how I feel!!! We should have all of our rights and freedoms to choose what is best for us. #freedom #ourbodyourchoice #NoVaccineForMe #novaccinepassport.
MRE	fetal/fetus/fetuses, Mark of the beast, guinea pig(s), lab rat(s), DNA, mRNA, medical ethics	(1) Vatican says use of Covid vaccines made from aborted fetal tissue is ethical. (2) Africans let's rise up and put an end to this menace.. We are not lab rats!! We are not test tubes!! #Nomorevaccinetesting
AnimalVac	#animalhealth, #animalwelfare, #WorldAnimalVaccinationDay, #petmedicine, #vetmedicine, Feline Panleukopenia, Feline Herpesvirus, Feline Calicivirus, Feline Leukaemia Virus, Canine Distemper Virus, Canine Parvovirus, Canine Adenovirus, Canine Rabies	(1) Will Your Pet Need a COVID-19 Vaccine? #covid19 #AnimalHealth (2) Outbreaks of disease are unpredictable and can have a major financial impact on your farm business. Vaccination is a planned approach to help to protect your livestock and improve animal health #VaccinesWork #WorldAnimalVaccinationDay

Table 3: Keywords and hashtags for data augmentation.

fier. The other VAE takes input as the concatenation of the BERT representation and the predicted label of the classifier. The output of the decoders is the bag-of-words of the input text. To evaluate the benefit of pre-training with in-domain data (Gururangan et al., 2020), we also experiment with a new variant – CANTM-COVID – where we replace BERT by COVID-Twitter-BERT (Müller et al., 2020) that is pretrained on COVID-19 related tweets.

BERT and BERT-COVID (Devlin et al., 2019; Müller et al., 2020): We fine-tune BERT (Devlin et al., 2019) and COVID-Twitter-BERT (Müller et al., 2020) model implemented on Hugging Face (Wolf et al., 2020) and follow the suggestion by Song et al. (2021) to enable a fair comparison between BERT and CANTM: an additional 500 dimensional feed-forward network is built on top of BERT and the parameters, except for BERT's last layer, are fixed during training.

5 Experimental Setup

5.1 Pre-processing and Hyperparameters

All user mentions, URLs, hashtags (including those we use for data augmentation) and emojis are removed from the posts. We use the suggested settings from the original implementations (Song et al., 2021; Pedregosa et al., 2011; Card et al., 2018) except for the following hyperparameters. For each hyperparameter tuning experiment, we randomly designated 20% of the data points in the training set as a development set. All possible combinations of candidate parameter values were tested and the optimal value was determined based on maximising the macro-F1 score on the development set.

For BERT, BERT-COVID, CANTM and CANTM-COVID, the batch size is searched from $\{16, 32, 64\}$. Since FD data contains posts with long textual length, we experiment with three truncation strategies (Sun et al., 2019): keep the beginning (the first 300, 400, or 512 tokens), the end (the last 300, 400, or 512 tokens) or a combination of both strategies (the first 300 and the last 212 tokens). The optimal selection in each experiment is keeping the first 400 tokens and training with batch size as 32. The same truncated texts are used for BOW-LR and SCHOLAR. For SCHOLAR, we set embedding dimension as 500, chosen from $\{300, 400, 500, 600\}$.

5.2 Evaluation

We compare the models based on 5-fold stratified cross validation on the augmented seven-class dataset. The average of macro-F1, accuracy and per-class F1 scores are reported.

6 Results

Table 4 presents the results of model comparison. The pre-trained transformer-based models significantly outperform BOW-LR and SCHOLAR whose model structures are much simpler. CANTM shows an increase in accuracy and macro-F1 scores compared with the strong baseline model BERT. Taking advantage of pre-training on an in-domain corpus of COVID tweets with a larger transformer model, BERT-COVID outperforms

CANTM. CANTM-COVID further improves the performance, achieving the highest accuracy and macro-F1 scores. Models tend to perform better on the Cons, LF, MRE and AnimalVac classes. This is expected, since they consist of posts retrieved through class-associated keywords.

7 Analysis

7.1 Evaluation of data augmentation

We analyse (1) whether our newly collected posts improve the performance on the minority classes in FD data; (2) whether the introduction of the AnimalVac class impacts the performance on the six human-related vaccine classes.

Data Split For the first purpose, we construct two training sets (Training set(imbalanced) and Training set(balanced)) and a test set (Test set(six-class)). The data of the six topics except for MRE in the FD data is randomly split in the ratio of 7:3 in the case of Training set(imbalanced) and Test set(six-class). Since the MRE class only consists of seven posts in the FD dataset, we include them in the Test set(six-class) only. The newly collected MRE posts are randomly split in the same ratio as above to complete the Training set(imbalanced) and Test set(six-class). The Training set(balanced) is the combination of Training set(imbalanced) and the rest of the new posts we collected during data augmentation.
To contrast the performance before and after the introduction of the new category AnimalVac, we randomly split the data points in the AnimalVac class into two parts (7:3) and add them into Training set(balanced) and Test set(six-class) respectively, that is, Training set(seven-class) and Test set(seven-class). Table 5 presents the statistics of the training and test data.

Experimental Setup We use CANTM-COVID for this set of experiments as it is the best performing model as shown above. We run each experiment five times and report the average of macro-F1 and accuracy scores.

Results The results are presented in Table 6. We also show the confusion matrices in Fig 1.

Re-balancing the training set could increase accuracy by 3% and the macro-F1 score by 10%. The recall scores of the two target minority classes (LF and Cons) grow from 0.31 to 0.49 and from 0.04 to 0.36 respectively, while the performance of the other four classes are not significantly influenced. As for the MRE class, 43% of posts in FD data can be correctly predicted if training with only the newly collected tweets for this class, either in imbalanced, balanced six-class or seven-class setting. We observe that the model could accurately identify all the short tweets in LF after data augmentation. However, it is still hard for the model to correctly classify long posts. Details about this shortcoming are discussed in the next section.

Introducing the AnimalVac class does not strongly impact the performance on the other six categories about human vaccination, which are the more important classes for this task. The model could accurately recognise 98% of posts regarding animal vaccination, denoting that animal vaccine posts are easily distinguishable.

As shown in Fig 1, PE and SEN posts are easily mis-classified as DPA (16% and 25% respectively. It is also hard for the model to distinguish LF from SEN and PE. The model struggles most on classifying the narratives about conspiracies. Only 32% of them can be correctly tagged even after data augmentation. We discuss the potential reasons and provide examples in the next section.

Furthermore, the drop in performance as compared to the results in Table 4 indicates that it is relatively easier for the model to learn and identify the augmented data collected through class-associated keyword matching, but hard to generalise to unseen domains, especially for the Cons class. It should be noted that we intentionally involve conspiracy stories that are not in the FD dataset (only "nano" and "deep state" appear in one post respectively after pre-processing). The LF class is less impacted since 95% of new posts are collected through hashtags which are removed before training. However, our results still illustrate promising improvement in performance over the target topics, showing the ability of model generalisation.

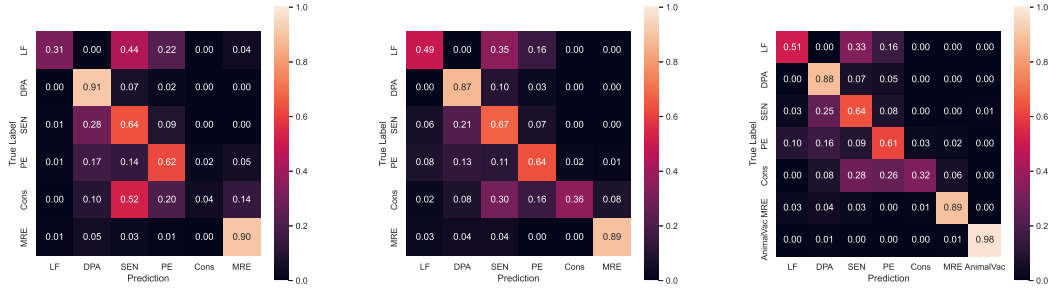
7.2 Error Analysis

Although our model performs well, we highlight the following challenges and limitations. We provide some error analysis examples in Table 7.

Text Length: Long narratives involving multiple topics are easily misclassified. As shown in Table 7, the first post cites safety considerations and side

Model	Macro-F1	Accuracy	F1 score						
			Cons	DPA	LF	MRE	PE	SEN	AnimalVac
BOW-LR	0.67	0.67	0.62	0.62	0.72	0.77	0.52	0.50	0.83
SCHOLAR	0.65	0.66	0.65	0.56	0.67	0.88	0.46	0.43	0.89
BERT	0.74	0.75	0.79	0.63	0.65	0.92	0.54	0.59	0.95
BERT-COVID	0.80	0.80	0.90	0.73	0.83	0.94	0.64	0.63	**0.97**
CANTM	0.77	0.77	0.82	0.70	0.75	0.94	0.60	0.62	0.96
CANTM-COVID	**0.84**	**0.84**	**0.91**	**0.77**	**0.86**	**0.96**	**0.67**	**0.72**	**0.97**

Table 4: Results of model performance on the augmented seven-class test dataset. The best results are in bold.



(a) (b) (c)

Figure 1: Confusion matrices for data augmentation evaluation. **(a)** Model trained on six-class imbalanced data. **(b)** Model trained on six-class re-balanced data. **(c)** Model trained on seven-class re-balanced data.

Datasets	**Cons**	DPA	**LF**	MRE	PE	SEN	**AnimalVac**
Training set (imbalanced)	16	81	26	114	76	94	0
Training set (balanced)	97	81	81	114	76	94	0
Test set (six-class)	10	35	11	37	32	40	0
Training set (seven-class)	97	81	81	114	76	94	67
Test set (seven-class)	10	35	11	37	32	40	29

Table 5: Label count of the training and test sets for the evaluation of data augmentation. The target classes are in bold.

Training set	Test set	Macro-F1	Accuracy
imbalanced	six-class	0.57	0.69
balanced	six-class	0.67	0.72
seven-class	seven-class	0.69	0.75

Table 6: Results of data augmentation evaluation of the CANTM-COVID model.

effects of vaccination as grounds for objecting to mandatory vaccination. In this case, the classifier incorrectly assigns the SEN label. The fourth claim shows another example whose true label is SEN while the model falsely tags it as DPA. The classifier is confused because the post elaborates on the development of the COVID-19 vaccine to support the opinion towards the necessity of the vaccine in the last sentence.

Temporal Drift: Dataset and model need to be updated over time, especially for the DPA and Cons classes, since new conspiracy theories are emerging continuously. The poor performance on the Cons class (see Fig 1b) illustrates that the model is finding it hard to generalise to new conspiracies. Also, progress concerning development, testing and provision of COVID-19 vaccination is fast changing. The samples in the DPA class were collected by First Draft in 2020 and most of the posts in their dataset refer to the announcement of the registration of the world's first COVID-19 vaccine by Russia, thus lacking examples of more recent events. Consequently we observe that the model tends to infer an unexpected correlation between Russian and the DPA class.

Model Bias: The size of the current dataset is still relatively small and this may result in model bias. As shown in the second example in Table 7, the mention of "Biden" and "Trump" may be the reason for the misclassification as they frequently appear in posts pertaining to politics. The class-associated words generated by CANTM-COVID confirm our assumption: "Trump" is highly associated with the PE class. Similarly, Bill Gates, who is often linked to conspiracy theories, is frequently involved in narratives about economics in

	True label	Prediction	Narrative
1	LF	SEN	....This is XXX - three months old, five days after a round of vaccines, showing the distinct sign of stroke. She died two days later....this type of asymmetry was common in the faces of the kids the day following vaccinations....Keep your eye, your focus on the MAIN GOAL: NO MANDATES period. No Mandates. No Mandates. Censorship is real.
2	LF	PE	Happy to be here after spending years suffering from Trump delusion syndrome....It seems the only policy Biden has spoken about is how he will mandate masks, which ultimately will lead to vaccine mandates. Biden is in the dark in terms of medical freedom. Trump for sure.
3	Cons	PE	We need to depopulate the planet. Also Bill Gates: Save your life with my vaccine.
4	SEN	DPA	Good News on Covid 19 vaccine: The result of the phase two trial of the Covid 19 vaccine by Oxford University's Jenner Institute and Oxford vaccine group is very positive. The result showed a strong immune response in both parts of the immune system. The vaccine provoked a T cell response within 14 days of vaccination that can attack cells infected with the Covid 19. Participants who received the vaccine also had detectable neutralising antibodies important for protection against Covid 19. Oh God, please make this vaccine work so that we can go back to our normal world. Amen/Ameen.

Table 7: Misclassification examples.

the training set. In fact, "Gates" is among the top 5 topics for the PE class, which may explain the misclassification of the 3rd conspiracy post. The class-associated keyword-based data augmentation may also make the model overly dependent on these target terms as discussed before.

8 Conclusion

This paper proposed a novel seven-way classification task for categorising online vaccine narratives. We augmented an existing six-class dataset semi-automatically, leading to a more balanced data distribution and the inclusion of an additional seventh category of posts related to animal vaccines. We experimented with strong baseline models and our best model CANTM-COVID achieves an accuracy score of 0.84 using 5-fold cross-validation. We also show that data augmentation of minority classes helps to produce better models, without significantly impacting the performance on the remaining classes. Moreover, the addition of the new animal vaccine category does not significantly influence model performance on the original six human vaccine related classes.

In our discussion, we highlighted the main challenges of this task and the current limitations of our model. Future work will focus on addressing some of those challenges, including development of models capable of dealing with longer posts.

Last but not least, our vaccine narratives classifier is made available through an API for reproducibility reasons. We believe this is a significant contribution towards understanding and tracking online debates around vaccine safety and hesitancy.

Acknowledgments

This research is supported by a University of Sheffield QR SPF Grant, an EPSRC research grant (EP/W011212/1 XAIvsDisinfo: eXplainable AI Methods for Categorisation and Analysis of COVID-19 Vaccine Disinformation and Online Debates), and an European Union Horizon 2020 Project (Agreement no.871042 under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities": "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu)). We would like to thank First Draft for data and codebook sharing and valuable feedback.

References

Rodrigo Agerri, Roberto Centeno, María Espinosa, Joseba Fernandez de Landa, and Alvaro Rodrigo. 2021. Vaxxstance@ iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, 67:173–181.

Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, et al. 2021. Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms. In *ICWSM*, pages 913–922.

Young Anna Argyris, Kafui Monu, Pang-Ning Tan, Colton Aarts, Fan Jiang, and Kaleigh Anne Wiseley. 2021. Using machine learning to compare provaccine and antivaccine discourse among the public on social media: Algorithm development study. *JMIR public health and surveillance*, 7(6):e23105.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Erika Bonnevie, Allison Gallegos-Jeffrey, Jaclyn Goldbarg, Brian Byrd, and Joseph Smyser. 2021. Quantifying the rise of vaccine opposition on twitter during the covid-19 pandemic. *Journal of communication in healthcare*, 14(1):12–19.

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040, Melbourne, Australia. Association for Computational Linguistics.

Mingxuan Chen, Xinqiao Chu, and KP Subbalakshmi. 2021. Mmcovar: multimodal covid-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 31–38.

Matthew R DeVerna, Francesco Pierri, Bao Tran Truong, John Bollenbacher, David Axelrod, Niklas Loynes, Christopher Torres-Lugo, Kai-Cheng Yang, Filippo Menczer, and John Bryden. 2021. Covaxxy: A collection of english-language twitter posts about covid-19 vaccines. In *ICWSM*, pages 992–999.

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.

Juhani Eskola, Philippe Duclos, Melanie Schuster, Noni E MacDonald, et al. 2015. How to deal with vaccine hesitancy? *Vaccine*, 33(34):4215–4217.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public health*, 203:23–30.

Brian Hughes, Cynthia Miller-Idriss, Rachael Piltch-Loeb, Beth Goldberg, Kesa White, Meili Criezis, and Elena Savoia. 2021. Development of a codebook of online anti-vaccination rhetoric to manage covid-19 vaccine misinformation. *International journal of environmental research and public health*, 18(14):7556.

Amelia M Jamison, David A Broniatowski, Mark Dredze, Anu Sangraula, Michael C Smith, and Sandra C Quinn. 2020. Not just conspiracy theories: Vaccine opponents and proponents add to the covid-19 'infodemic'on twitter. *Harvard Kennedy School Misinformation Review*, 1(3).

Florian Kunneman, Mattijs Lambooij, Albert Wong, Antal van den Bosch, and Liesbeth Mollema. 2020. Monitoring stance towards vaccination in twitter messages. *BMC medical informatics and decision making*, 20(1):1–14.

Heidi J Larson, Caitlin Jarrett, Elisabeth Eckersberger, David MD Smith, and Pauline Paterson. 2014. Understanding vaccine hesitancy around vaccines and vaccination from a global perspective: a systematic review of published literature, 2007–2012. *Vaccine*, 32(19):2150–2159.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Toward a multilingual and multimodal data repository for covid-19 disinformation. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4325–4330. IEEE.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Joanne Chen Lyu, Eileen Le Han, and Garving K Luli. 2021. Covid-19 vaccine–related discussion on twitter: topic modeling and sentiment analysis. *Journal of medical Internet research*, 23(6):e24435.

Yida Mu, Mali Jin, Charlie Grimshaw, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter. *arXiv preprint arXiv:2301.06660*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Goran Muric, Yusong Wu, Emilio Ferrara, et al. 2021. Covid-19 vaccine hesitancy on social media: building a public twitter data set of antivaccine content, vaccine misinformation, and conspiracies. *JMIR public health and surveillance*, 7(11):e30642.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Iftikhar Qayum. 2019. Top ten global health threats for 2019: the who list. *Journal of Rehman Medical Institute*, 5(2):01–02.

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv:2006.11343*.

Karishma Sharma, Yizhou Zhang, and Yan Liu. 2022. Covid-19 vaccine misinformation campaigns and social media narratives. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 920–931.

Junaid Shuja, Eisa Alanazi, Waleed Alasmary, and Abdulaziz Alashaikh. 2021. Covid-19 open source data sets: a comprehensive survey. *Applied Intelligence*, 51(3):1296–1325.

R Smith, S Cubbon, and C Wardle. 2020. Under the surface: Covid-19 vaccine narratives, misinformation & data deficits on social media. *USA: First Draft*.

Xingyi Song, Johann Petrak, Ye Jiang, Iknoor Singh, Diana Maynard, and Kalina Bontcheva. 2021. Classification aware neural topic model for covid-19 disinformation categorisation. *PloS one*, 16(2):e0247086.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jia Xue, Junxiang Chen, Ran Hu, Chen Chen, Chengda Zheng, Yue Su, Tingshao Zhu, et al. 2020. Twitter discussions and emotions about the covid-19 pandemic: Machine learning approach. *Journal of medical Internet research*, 22(11):e20550.

Koosha Zarei, Reza Farahbakhsh, Noel Crespi, and Gareth Tyson. 2020. A first instagram dataset on covid-19. *arXiv preprint arXiv:2004.12226*.