Examining Temporalities on Stance Detection Towards COVID-19 Vaccination

Yida Mu, Mali Jin, Kalina Bontcheva, Xingyi Song
Department of Computer Science, The University of Sheffield
{y.mu, m.jin, k.bontcheva, x.song}@sheffield.ac.uk

Abstract

Previous studies have highlighted the importance of vaccination as an effective strategy to control the transmission of the COVID-19 virus. It is crucial for policymakers to have a comprehensive understanding of the public's stance towards vaccination on a large scale. However, attitudes towards COVID-19 vaccination, such as pro-vaccine or vaccine hesitancy, have evolved over time on social media. Thus, it is necessary to account for possible temporal shifts when analysing these stances. This study aims to examine the impact of temporal concept drift on stance detection towards COVID-19 vaccination on Twitter. To this end, we evaluate a range of transformer-based models using chronological (splitting the training, validation, and test sets in order of time) and random splits (randomly splitting these three sets) of social media data. Our findings reveal significant discrepancies in model performance between random and chronological splits in several existing COVID-19-related datasets; specifically, chronological splits significantly reduce the accuracy of stance classification. Therefore, real-world stance detection approaches need to be further refined to incorporate temporal factors as a key consideration.

Keywords: Stance Detection, COVID-19, Vaccine Hesitancy, Temporal Concept Drift

1. Introduction

The COVID-19 pandemic has had a profound impact on global health and has resulted in considerable social and economic disruption (Ciotti et al., 2020). Promoting vaccination has been statistically recognised as a vital tactic in curbing the spread of the COVID-19 virus and reducing the burden on healthcare systems (Lopez Bernal et al., 2021). However, attitudes towards COVID-19 vaccination have varied, with some individuals expressing hesitation and resistance (Cotfas et al., 2021; Poddar et al., 2022a). These concerns are caused by factors including side effects, conspiracy theories, and distrust of healthcare authorities (Poddar et al., 2022b). To promote vaccine uptake, it is important to understand the factors that contribute to hesitant or negative attitudes towards COVID-19 vaccination on a large scale (Mu et al., 2023b). Recently, there has been a growing interest in using supervised machine learning approaches to automatically detect users' stance towards COVID-19 vaccination on social media (Di Giovanni et al., 2022; Glandt et al., 2021; Chen et al., 2022).

Temporal concept drift refers to the phenomenon in NLP where the statistical properties of a dataset change over time such as the distribution of topics (Huang and Paul, 2019). It is particularly relevant in applications where data is collected over extended periods, such as in financial forecasting or tasks based on social media data (Xing et al., 2018; Alkhalifa et al., 2023; Hu et al., 2023). It can also lead to the degradation of model performance due to the temporal variation of textual content in a static dataset (Mu et al., 2023a,c). The impact of temporal concept drift has been investigated in several domains including topic classification (Chalkidis and Søgaard, 2022), rumour detection (Mu et al., 2023a), gender equality (Alkhalifa et al., 2021), and hate speech detection (Florio et al., 2020; Jin et al., 2023). Moreover, (Alkhalifa et al., 2023) evaluated how different factors of datasets and models affect the performance over time across different classification tasks.

However, temporal aspects have not been studied in stance detection concerning COVID-19 vaccination. Furthermore, the underlying distribution of users' stances may change over time due to factors such as evolving political agendas and emerging viral variants, making it necessary to take the temporal factor into account.

This study examines the temporal concept drift in stance detection towards COVID-19 vaccination on Twitter for the first time.[1] Specifically, we focus on the following research questions:

- ***Q1:*** Does temporal concept drift exert a significant affect on COVID-19 vaccine stance detection, as previously seen in other domains (e.g., rumour, legal, biomedical, etc.)?
- ***Q2:*** How does the model's performance differ across multiple languages?
- ***Q3:*** Can domain adaptation approaches be employed to mitigate the temporal concept drift impact on stance classification?

[1] Our code: `https://github.com/YIDAMU/COVID_Temporalites`

LREC-COLING 2024, pages 6732–6738

Dataset	Time	Tweets	Labels	Language
Cotfas et al. (2021)	Nov 2020 ∼Dec 2020	2,792	in favour, against, neutral	en
Poddar et al. (2022a)	Jan 2020 ∼March 2021	1,700	in favour, against, neutral	en
Mu et al. (2023b)	Nov 2020 ∼April 2022	3,101	pro, anti, hesitancy, irrelevant	en
Chen et al. (2022)	Jan 2020 ∼March 2021	17,934	pos, neg, neutral, off-topic	fr, de, en
Di Giovanni et al. (2022)	Nov 2020 ∼ June 2021	3,101	in favour, against, neutral	es, de, it

Table 1: Dataset statistics.

- ***Q4:*** Does the semantic variation between the training and test sets lead to a degradation or improvement of the model's predictive performance?

To achieve this, we (i) evaluate five publicly available monolingual and multilingual datasets, (ii) conduct a set of controlled experiments by evaluating various transformer-based pretrained language models (PLMs) using *chronological* and *random splits* and (iii) perform correlation tests to examine the relationship (i.e., positive or negative) between the model predictive performance and the disparity between the two subsets (i.e., training and test sets).

2. Experimental Setup

2.1. Datasets

We use three datasets [2] in English (Cotfas et al., 2021; Poddar et al., 2022a; Mu et al., 2023b) and two datasets[3] in multiple languages (Chen et al., 2022; Di Giovanni et al., 2022). These datasets adhere to the FAIR principles (i.e., Findable, Accessible, Interoperable and Re-usable) (Wilkinson et al., 2016). More details (e.g. sources, annotation details, label definitions) can be found in the original articles. Differences in specifications between these datasets are shown in Table 1.

2.2. Data Splits

We aim to conduct a set of controlled experiments to explore the impact of temporality on the accuracy of stance classifiers regarding COVID-19 vaccination. To this end, we evaluate two data split strategies:

- **Chronological Splits** Following Mu et al. (2023a), all datasets are sorted chronologically and subsequently divided into a training set (70% earliest data), a validation set (10% data after training set and before test set) and a test set (20% latest data). Note that the three subsets do not overlap temporally.

- **Random Splits** We randomly split all datasets using a stratified 5-fold cross-validation method, ensuring that the class proportions in each fold mirror those in the original dataset. Additionally, the ratio of training to validation to testing matches that of the chronological splits.

2.3. Models

We evaluate various transformer-based PLMs. Following Devlin et al. (2019), we fine-tune these PLMs by adding a fully-connected layer on top of the transformer architecture. We consider the special token '[CLS]' as the tweet-level representation.

Mono-lingual PLMs To represent tweets in English, we consider vanilla BERT and two domain-adapted PLMs:

- **BERT** (Devlin et al., 2019) is trained on a large corpus including Wikipedia articles and English Books-Corpus.

- **COVID-BERT** (Müller et al., 2023) is a specialised version of BERT model that has been further pre-trained on a large corpus of COVID-19 related texts to improve the performance of related downstream tasks (Cotfas et al., 2021; Poddar et al., 2022a).[4]

- **Vaccine-BERT** (Mu et al., 2023b) is a domain adapted BERT model for automatically detecting COVID-19 vaccine stance.[5] It has been further pre-trained on a large dataset of tweets related to *COVID-19 vaccines* based on the COVID-BERT model to improve its ability to accurately classify the vaccine stance.

Multilingual PLMs To represent tweets in multilingual, we consider two strong cross-lingual PLMs:

[2](i) https://github.com/liviucotfas/covid-19-vaccination-stance-detection; (ii) https://github.com/sohampoddar26/covid-vax-stance; (iii) https://zenodo.org/records/7601328

[3](iv) https://zenodo.org/records/5851407; (v) https://github.com/datasciencepolimi/vaccineu

[4]https://huggingface.co/digitalepidemiologylab/covid-twitter-bert-v2

[5]https://huggingface.co/GateNLP/covid-vaccine-twitter-bert

Model	Splits	Cotfas et al. (2021)				Poddar et al. (2022a)				Mu et al. (2023b)			
		P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
BERT	*Random*	79.2	79.3	79.2	79.4	50.8	47.6	43.5	58.4	54.6	54.6	54.4	54.5
	Chronological	62.9	64.6	63.3	74.5	46.9	45.4	41.1	57.4	52.9	53.2	52.5	53.0
COVID BERT	*Random*	85.8	85.1	85.3	85	71.2	68.6	69.3	71.0	67.3	67.2	67.2	67.2
	Chronological	74.2	75	**74.5**	80.4	67.7	68.0	**67.6**	70.5	67.6	68.0	**67.7**	68.3
VAXX BERT	*Random*	85.4	85.3	85.4	85.1	71.2	66.4	67.5	69.9	67.7	67.7	67.6	67.6
	Chronological	71.8	71.2	70.7	79.3	65.5	63.7	64.1	66.8	67.5	67.3	67.4	68.2

Table 2: Model predictive performance on mono-lingual datasets. Cells in *Grey* denotes that the classifier trained on random splits performs significantly better than chronological splits ($p < 0.05$, t-test). The smallest performance drop (or increase) using chronological splits is in bold.

Model	Splits	Di Giovanni et al. (2022)				Chen et al. (2022)			
		P	R	F1	Acc	P	R	F1	Acc
XML-BERT	*Random Splits*	42.2	41.6	41.6	52.8	63.5	62.1	62.7	75.4
	Chronological Splits	42.4	41.8	**42.0**	52.9	60.6	60.6	57.9	73.4
XML-RoBERTa	*Random Splits*	45.8	43.9	44.2	55.1	65.0	64.0	64.5	77.6
	Chronological Splits	45.2	43.4	43.7	55.0	62.4	61.9	**62**	75.5

Table 3: Model predictive performance on multilingual datasets. Cells in *Grey* denotes that the classifier trained on random splits performs significantly better than chronological splits ($p < 0.05$, t-test). The smallest performance drop (or increase) using chronological splits is in bold.

- **XLM-BERT** (Devlin et al., 2019) is a multilingual version of BERT that has been pre-trained on texts from over 100 multiple languages.
- **XLM-RoBERTa** (Conneau et al., 2020) is trained to reconstruct a sentence in one language from a corrupted version of the sentence in another language, which has been shown highly effective for multilingual NLP tasks such as cross-lingual stance classification (Chen et al., 2022).

2.4. Training & Evaluation

For all datasets, we keep the original setup (i.e., multi-class classification task). We pre-process the tweets from all datasets by (i) lowercasing and (ii) replacing @user_name and hyperlinks with special tokens i.e., '@USER' and 'HTTPURL' respectively.

All models are trained on the training set, while model tuning and selection are based on the validation loss observed at each training epoch. Subsequently, the predictive performance of the model is assessed on the test set. We run all models five times with varying random seeds to ensure consistency and report the average Accuracy, Precision, Recall, and macro-F1 scores. For all PLMs, we set learning rate as 2e-5, batch size as 16, and max number of input tokens as 256. All experiments are performed on a NVIDIA Titan RTX GPU with 24 GB memory.

3. Analysis

In this section, we present a detailed similarity analysis at different levels (e.g., token and topic) and conduct an error analysis based on the output of the model.

3.1. Text Similarity

Our aim is to investigate whether a decrease in model predictive performance occurs due to variations between the two subsets used for training and testing, and whether the difference in performance lessens as the datasets become more similar to each other. Following Kochkina et al. (2023); Jin et al. (2023), we measure the difference between training and test sets for chronological and random splits using two matrices: (i) Intersection over Union (IoU); (ii) DICE coefficient (DICE) (Dice, 1945).

Intersection over Union

$$IoU = \frac{|V^p \cap V^q|}{|V^p \cup V^q|} \tag{1}$$

DICE coefficient

$$DICE = \frac{2 \times |V^p \cap V^q|}{|V^p| + |V^q|} \tag{2}$$

where V^p and V^q denote the lists of unique tokens from two subsets (i.e., training and test sets) respectively. $|V^p \cap V^q|$ and $|V^p \cup V^q|$ denote the total number of unique tokens that appear in the **intersection** and **union** of the two subsets respectively. When the two subsets have no shared vocabulary, the IoU and DICE values will be zero, while if they are identical, the IoU and DICE values will be equal to 1.

3.2. Topics Drift

We also employ BERTopic (Grootendorst, 2022) to examine the temporal concept drift at the topic

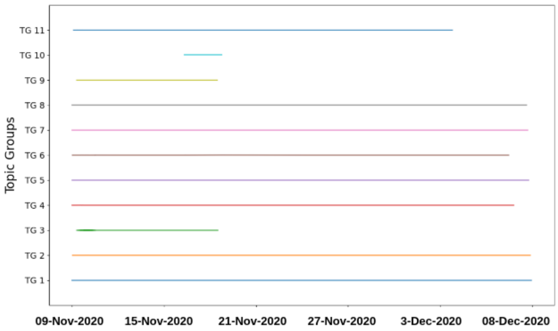

Figure 1: Topic distribution over time from Cotfas et al. (2021). TG is short for 'topic group'.

level on the Cotfas dataset (Cotfas et al., 2021). We first extract the top 15 topic groups in the dataset using BERTopic. Then we manually combine similar topic groups and delete repeated topics and commonly used words (e.g., *you*). Figure 1 displays the distribution of 11 topic groups across the entire dataset over time. We observe that some topics only occur during certain periods, which indicates that using temporal data splits may yield an imbalanced topic distribution between the training and test sets. The phenomenon of topic drift may lead to a drop in model performance when using chronological data splits.

3.3. Error Analysis

We also manually conduct an analysis to investigate the model behaviours with two splitting strategies. First, when using random splits, we observe that some tweets in the test set are similar or identical to tweets in the training set using random splits rather than chronological splits (similar tweets are more likely to be generated during the same time period). This leads to a higher prediction accuracy when using random splits. For example, two pairs of tweets from Chen et al. (2022) and Cotfas et al. (2021) are identical after data pre-processing:

> **Tweet_id:1220414** & Tweet_id:1220415**** *This deadly #coronavirus was spread fr #wuhan (China) to many countries just in days and kill 17 ppl. Maybe thousand ppl carrying pathogens are traveling around the worldIts spread fr human to human via air and still have no vaccinesPls read it; reduce the risk of infection. HTTPURL*

> **Tweet_id:133460** & Tweet_id:133520**:** *The U.S. #airline industry and its pilots are essential to the distribution of a COVID-19 vaccine. Congress and government leaders must #ExtendPSP now to ensure critical infrastructure is in place to distribute a vaccine—American lives depend on it @USER*

Datasets	Splits	IoU	DICE	Acc
Cotfas	*Random*	0.17	0.25	85.1
	Chronological	0.14	0.22	79.3
Poddar	*Random*	0.16	0.18	71.0
	Chronological	0.14	0.17	70.5
Mu	*Random*	0.22	0.27	67.6
	Chronological	0.22	0.26	68.3
Di	*Random*	0.12	0.14	55.1
	Chronological	0.11	0.13	55.0
Chen	*Random*	0.18	0.21	77.6
	Chronological	0.17	0.20	75.5
Pearson coefficient		**0.35**	**0.64**	-

Table 4: IoU and DICE values between training and test sets. **Acc** represents the best model accuracy. We also display the Pearson correlation between the two values and best accuracy values across all models.

Also, we observe that some tweets containing emerging topics (topics that appear in the later time period only) are correctly classified using random splits (topics overlap in both training and test sets) but wrongly using chronological splits. An example from Mu et al. (2023b) is shown below.

> **Tweet_id:148891****: *@USER* ***Vaccine passes*** *also impede freedom of movement. The irony.* 🤡🤡🤡

The data from Mu et al. (2023b) covers from Nov 2020 to April 2022 while 'Vaccine passes' was introduced in May 2021[6]. It is likely to cause models to fail to identify it in the testing set as models are unable to learn from the training set using chronological splits.

4. Discussion

Q1 & Q2: Chronological vs Random Splits in Multiple Languages In general, we notice that using random splits leads to an overestimation of performance compared to using chronological splits across the majority PLMs. Our findings align with the prior studies on temporal concept drift (Chalkidis and Søgaard, 2022; Mu et al., 2023a). However, previous work has shown the stance detection results are vulnerable to simple perturbations (Schiller et al., 2021), which explains the results are not consistent over datasets from Mu et al. (2023b) and Di Giovanni et al. (2022) (the performance increases using chronological splits). Furthermore, we observe similar model performance for both data splitting strategies on Mu et al. (2023b) and Di Giovanni et al. (2022) datasets. Note that the results of the two distance measures (i.e., IoU and DICE) between the training and test set are also similar (Table 4).

[6] https://www.instituteforgovernment.org.uk/article/explainer/covid-passports

Q3: Vanilla vs. Domain Adapted PLMs We refer to the decrease in F1-score using chronological splits versus random splits as *performance drop*. For all mono-lingual datasets, we observe that the performance drops less using domain-adapted PLMs (i.e., COVID-BERT and VAXX-BERT) than using vanilla BERT models. Taking COVID-BERT model for example, F1 scores decrease 10.8% (-15.9% for BERT), 1.7% (-2.4% for BERT) and even increase 0.5% (-1.9% for BERT) for datasets of Cotfas et al. (2021), Poddar et al. (2022a) and Mu et al. (2023b) respectively (see Table 2). This indicates that applying domain & task adaptation techniques can address the issue of temporal concept drift to a certain extent in stance detection towards COVID-19 related datasets. We also notice that f1 scores drop less using COVID-BERT than VAXX-BERT (e.g., -10.8% vs. -14.7% on the data set of Cotfas et al. (2021)). We speculate that this is because a more domain-specific model (i.e., VAXX-BERT) lead to poorer generalise ability and is less sensitive to time.

Q4: Distance Between Training and Test Sets In table 4, we observed that using random splits results in significantly higher IoU and DICE scores (note that higher scores indicate greater similarities between the training and test sets) compared to chronological splits. This suggests that new topics (i.e., temporal concept drift) emerge in the test sets when using the chronological split strategy. Also, we discover a positive Pearson correlation between the model accuracy and the similarity distance of two subsets using both IoU (0.35) and DICE (0.64) metrics, i.e., the higher the values, the higher the model accuracy.

5. Conclusion

We explored how temporalities affect stance detection towards COVID-19 vaccination on Twitter. Our experiments showed that using chronological splits significantly reduces the accuracy of stance classification in existing datasets. Therefore, we believe that developing real-world stance detection approaches should take temporal factors into account. Meanwhile, our results suggest that using domain- and task-adaptive models, and combining models trained on different time periods, can effectively address the effects of temporal concept drift in COVID-19 vaccination stance detection.

Ethics Statement

Our work has been approved by the Research Ethics Committee of our institute, and complies with the policies of Twitter API. All datasets are publicly available via the links (see footnotes) provided in the original papers.

Acknowledgements

This research is supported by a UKRI grant EP/W011212/1 ("XAIvsDisinfo: eXplainable AI Methods for Categorisation and Analysis of COVID-19 Vaccine Disinformation and Online Debates")[7] and an EU Horizon 2020 grant (agreement no.871042) ("So-BigData++: European Integrated Infrastructure for Social Mining and BigData Analytics")[8].

References

Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. 2021. Opinions are made to be changed: Temporally adaptive stance classification. In *Proceedings of the 2021 workshop on open challenges in online social networks*, pages 27–32.

Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. 2023. Building for tomorrow: Assessing the temporal persistence of text classifiers. *Information Processing & Management*, 60(2):103200.

Ilias Chalkidis and Anders Søgaard. 2022. Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2441–2454.

Ninghan Chen, Xihui Chen, and Jun Pang. 2022. A multilingual dataset of covid-19 vaccination attitudes on twitter. *Data in Brief*, 44:108503.

Marco Ciotti, Massimo Ciccozzi, Alessandro Terrinoni, Wen-Can Jiang, Cheng-Bin Wang, and Sergio Bernardini. 2020. The covid-19 pandemic. *Critical reviews in clinical laboratory Sci.*, 57(6):365–388.

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

Liviu-Adrian Cotfas, Camelia Delcea, Ioan Roxin, Corina Ioanăş, Dana Simona Gherai, and Federico Tajariol. 2021. The longest month: analyzing covid-19 vaccination opinions dynamics

[7] `https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/W011212/1`

[8] `http://www.sobigdata.eu`

from tweets in the month following the first vaccine announcement. *Ieee Access*, 9:33203–33223.

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.

Marco Di Giovanni, Francesco Pierri, Christopher Torres-Lugo, and Marco Brambilla. 2022. Vaccineu: Covid-19 vaccine conversations on twitter in french, german and italian. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1236–1244.

Lee R Dice. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12):4180.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023. Learn over past, evolve for future: Forecasting temporal trends for fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 116–125, Toronto, Canada. Association for Computational Linguistics.

Xiaolei Huang and Michael Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123.

Mali Jin, Yida Mu, Diana Maynard, and Kalina Bontcheva. 2023. Examining temporal bias in abusive language detection. *arXiv preprint arXiv:2309.14146*.

Elena Kochkina, Tamanna Hossain, Robert L Logan IV, Miguel Arana-Catania, Rob Procter, Arkaitz Zubiaga, Sameer Singh, Yulan He, and Maria Liakata. 2023. Evaluating the generalisability of neural rumour verification models. *Information Processing & Management*, 60(1):103116.

Jamie Lopez Bernal, Nick Andrews, Charlotte Gower, Eileen Gallagher, Ruth Simmons, Simon Thelwall, Julia Stowe, Elise Tessier, Natalie Groves, Gavin Dabrera, et al. 2021. Effectiveness of covid-19 vaccines against the b. 1.617. 2 (delta) variant. *New England Journal of Medicine*, 385(7):585–594.

Yida Mu, Kalina Bontcheva, and Nikolaos Aletras. 2023a. It's about time: Rethinking evaluation on rumor detection benchmarks using chronological splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 724–731.

Yida Mu, Mali Jin, Charlie Grimshaw, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023b. Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1052–1062.

Yida Mu, Xingyi Song, Kalina Bontcheva, and Nikolaos Aletras. 2023c. Examining the limitations of computational rumor detection models trained on static datasets. *arXiv preprint arXiv:2309.11576*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2023. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence*, 6:1023281.

Soham Poddar, Mainack Mondal, Janardan Misra, Niloy Ganguly, and Saptarshi Ghosh. 2022a. Winds of change: Impact of covid-19 on vaccine-related opinions of twitter users. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 782–793.

Soham Poddar, Azlaan Mustafa Samad, Rajdeep Mukherjee, Niloy Ganguly, and Saptarshi Ghosh. 2022b. Caves: A dataset to facilitate explainable classification and summarization of concerns towards covid vaccines. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3154–3164.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance Detection Benchmark: How Robust is Your Stance Detection? *KI-Künstliche Intelligenz*, pages 1–13.

Mark D Wilkinson, Michel Dumontier, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Frank Z Xing, Erik Cambria, and Roy E Welsch. 2018. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73.