The Role of Context in Vaccine Stance Prediction for Twitter Users

Aleney Khoo[1,2] **Maciej Rybinski**[1] **Sarvnaz Karimi**[1] **Adam Dunn**[2]
[1]CSIRO Data61, Sydney, Australia
[2]The University of Sydney, Sydney Australia
{firstname.lastname}@csiro.au
adam.dunn@sydney.edu.au

Abstract

Public expression of vaccine related sentiment on social media platforms can be used in information surveillance applications to gain insight into vaccine hesitancy and its spread. Effective identification of vaccine-negative content constitutes one of the most fundamental building blocks in such applications.

Here, we investigate the role of users' previous vaccine-related posts, in the capacity of stance text classifiers to detect vaccine-negative content. We conduct experiments on a dataset of over 7K tweets manually labeled for vaccination stance captured between 2017 and 2019, with unlabeled historical data.

Our results indicate that incorporating user-generated-context improves stance detection. It also bridges the effectiveness gap between simple linear models and state-of-the art text classifiers, highlighting the importance of data capture strategy to the downstream task.

1 Introduction

Vaccine hesitancy is defined as delayed acceptance or refusal of vaccines despite their availability. It is believed that vaccine hesitancy and refusal may be amplified by the spread of vaccine-negative content on social media (Raballo et al., 2022).

Stance detection is a process of identification of speaker's judgment or standpoint towards a given proposition (Biber and Finegan, 1988; Mohammad et al., 2017). It is often modelled as a classification task of assigning an *against*, *for* or *neural* label to a text for a given *target*. In the context of social media, detecting anti-vaccine content can be seen as a building block fundamental to implementation of mitigation and monitoring strategies, understanding fears and concerns expressed in the public discourse (Mitra et al., 2021). In this problem, vaccine hesitancy is considered the target for stance detection.

In this paper we simplify the problem of stance detection by modelling it as a binary text classification task. In this set-up we investigate the impact of including users' historic tweets in detecting vaccine-negative utterances. We conduct experiments on a dataset of over 7K tweets manually labelled for vaccine stance captured between 2017 and 2019, which contains unlabelled historical data from the authors of the labelled tweets. Our work focuses on the impact of the availability of historic user-generated content on effectiveness on the downstream text classification task. We compare Transformer-based models, which allow for modelling the historical context in a cross-encoder, with traditional linear models incorporating these historical tweets in a Bag-of-Words (BoW) representation encompassing the labelled utterance.

2 Related Work

Our study relates to the literature on using social media for public health and, in particular, for detection of vaccination hesitancy. It also relates to Twitter, and other social media, text classification, sentiment and stance detection.

Social Media for Public Health Natural Language Processing (NLP) techniques for social media have been leveraged in different public health applications (Paul and Dredze, 2017; Conway et al., 2019), including for mental health (Calvo et al., 2017), syndromic surveillance (Jimeno Yepes et al., 2015; Ofoghi et al., 2016; Huang et al., 2016) of acute diseases (Joshi et al., 2020b) or infectious diseases (Joshi et al., 2019, 2020a), detecting user behaviour towards vaccination (Joshi et al., 2018), and personal health mention detection (Iyer et al., 2019). A survey of different social media platforms utilised for public health is presented by (Conway et al., 2019), showing a range of different platforms such as Twitter, Whatsapp, Facebook, and Reddit, as well as applications and methods.

Vaccination hesitancy detection Social media, in particular, Twitter have been a data source for gauging the public opinions on vaccines. Morante et al. (2020) present the Vaccination Corpus which annotated a corpus of 294 online debates published in news, blogs, editorial, governmental reports, science articles for their stance towards measles vaccines. Lanyi et al. (2022) analysed Twitter for COVID-19 vaccine hesitancy in order to identify barriers to vaccination in the UK. They approached this as a sentiment analysis and then mapped the tweets to predefined set of potential barriers such as safety or mistrust.

Mitra et al. (2021) studied Twitter data for a four-year period to understand anti-Vaccination attitudes. They used tweets of different users as their context to identify whether they are anti-vaccination. For their analysis they used topic modelling.

Stance Detection in social media Stance and sentiment detection on social media text, especially tweets, pose difficulties. Tweets are short and it can be difficult to identify the user's view either in terms on sentiment (positive, negative, neural) or stance (pro, anti, neutral) on a specific topic (Mohammad et al., 2017).

Medford et al. (2020) emphasise on the importance of Twitter data sentiment analysis during an outbreak of an infectious disease. They processed a large set of tweets using sentiment analysis and topic modelling in the early stages of the COVID-19 pandemic to help understand the effect of the outbreak on the public's emotions and beliefs.

Conforti et al. (2020) annotated a large corpus of tweets (over 51 thousand) for stance detection. The dataset represents public expressions of opinion on mergers and acquisition operations between companies. They also benchmark a number of different stance detection methods, including traditional ones such as SVM and those based on neural networks, such as CrossNet (Zheng et al., 2018). Conforti et al. (2021) uses this dataset to investigate cross-domain learning for stance detection when annotated data does not exist for a given target.

Stance detection using neural network-based methods is also investigated by Xu et al. (2018). They experimented with two different datasets from Twitter and showed promising results for cross-target stance detection using three methods based on BiLSTM (Zhou et al., 2016), MITRE (Augenstein et al., 2016).

Stance detection for opinions towards vaccination is studied by Skeppstedt et al. (2017). They annotated data from the British parental website Mumsnet for three labels of 'against', 'for', and 'undecided' and trained linear SVMs for stance detection.

Tweet Classification using Context Literature has long investigated the potential of context in classification of microblogs (or tweets) as a method to include more information to an otherwise short text. Historical tweets have been used in stance detection on Twitter, namely in fake news and sarcasm detection. Dou et al. (2021) used historical tweets to create a fake news detection framework that fuses historical tweets, news reports and engagement across user networks. Historical tweets can be incorporated in multiple way. Most commonly, they are bundled per user into one *document*. Chaudhry and Lease (2022) assess the impact of adding historical tweets in groups on a LSTM classifier, where the output is then fed into a Gradient Boosted Decision Tree classifier. They choose to retrieve up to 20 historical tweets per user, and separate groups of five. They highlighted the importance of context in these classification tasks, finding qualitatively that tweets were labelled often incorrectly on their own, but with context of a users historical tweets it could correctly label the tweet.

3 Dataset and Experimental Setup

The original dataset We conduct our experiments on a Twitter-based dataset by Dunn et al. (2020). The dataset consists of 10,080 vaccine-related Twitter posts (tweets) manually labelled for vaccine stance (anti-vaccine, pro-vaccine, other/neutral). The tweets were collected with vaccine-specific Twitter queries between January 12, 2017, and December 3, 2019 from U.S. based Twitter users with then-active accounts. The dataset also contains unlabelled historical vaccine-related (i.e., collected with the same queries) tweets from the authors of the labelled tweets. Each tweet in the dataset is represented with a user handle, timestamp, and the tweet content.

The task We frame the stance detection task in our experiments as a binary text classification problem with the focus on detecting the anti-vaccine content. The binarisation is, therefore, straightforward: we treat both the vaccine-positive and neutral classes as a new (binary) negative class, with the tweets labelled as vaccine-negative becoming the

new positive class.

Filtering The dataset comes with retweets filtered out (from both labelled and unlabelled data). Additionally, we filter the labelled data based on the availability of historical tweets – we only use tweets from users that have at least four historical tweets in the unlabelled portion of the dataset.

Preprocessing Since the focus of our experiments is on text-content-based stance detection in tweets, we set out to minimise the impact of network-specific features creeping into this textual content. We, therefore, normalise all user mentions with a 'USERNAME' placeholder.

Data at a glance Our dataset after filtering consists of 7,194 labelled tweets from 7,194 unique users (794 positive class/vaccine-negative and 6400 negative class/vaccine-positive-or-neutral), with every user in this dataset having at least 4 unlabelled historical tweets.

Experiments and setup We split the data into training and testing sets, with an 80–20 proportion. The training tweets were posted prior to the test tweets. In experiments where the historical tweets are incorporated, they are incorporated both at training and testing time. For traditional baselines (logistic regression–LR–and SVM) we incorporate the historical tweets by simply appending their text to the text of the labelled tweet (so, the historical context is modelled in the same BOW representation as the labelled tweet). For transformer-based models (RoBERTa variants) the historical tweets are appended after a SEP token (so, the labelled tweet becomes Sentence A of the BERT input, while the historical tweets are concatenated and fed as Sentence B part of the input). The hyperparameter tuning for LR and SVM was done with 3-fold cross-validation with grid search. For BERT-based models the hyperparameters for fine-tuning (batch size, number of epochs, learning rate) were tuned manually on a validation set (25% of the training data). This manual tuning was performed once for RoBERTa-base model with no historical tweets and its results (batch size of 16, learning rate of 2e-5, and 1 epoch of training) were applied directly to all other experiments with BERT derivatives. For each of the models we report results with no historical tweets, and with 1, 2, 3, and 4 historical tweets included in the training and inference.

We experiment with RoBERTa-base (henceforth referred to as 'plain RoBERTa') model as a domain agnostic BERT variant. We use a model trained for sentiment detection on Twitter[1] as a Twitter-optimised initial checkpoint. We chose RoBERTa variants over BERT due to more stable training and higher effectiveness in our initial experiments.

For the more successful of the two RoBERTa variants we run an additional experiment, where the predictions are produced only with the 4 historical context (so, the text of the actual training/test tweet is not used), to illustrate the predictive power behind the historical tweets. All RoBERTa results are averaged across 5 runs.

Where comparison are made between the results, we use approximate paired randomisation test to test for statistical significance of our findings. For transformer-based models we use the predictions resulting in median F1 score for significance testing, where not stated otherwise.

Dealing with imbalance While imbalanced classification adds a layer of complexity to our task, dealing with class imbalance is not our core focus. We therefore deal with the skew in our training dataset using standard approaches. In SVM and logistic regression we use regularisation inversely proportional to class size. In RoBERTa models we oversample the minority (vaccine-negative) class (10-fold). Both approaches yielded improvements of effectiveness on a validation set in our exploratory experiments, so we decided to incorporate them across the board.

4 Results

Our experiments on comparing different methods and different levels of context are presented in Table 1. We report precision, recall and F1-Score on the minority class (vaccine-negative). We observe the best results for plain RoBERTa with user-context incorporated by appending 4 historical tweets. Improvements in classification effectiveness can be seen across the board with incorporation of historical tweets, with linear models improving more, when compared to respective runs with no user-context.

5 Discussion

Our results demonstrate that adding historical tweets improves vaccine-negative stance detection

[1] cardiffnlp/twitter-roberta-base-sentiment-latest

Method	Precision	Recall	F1-Score
LogReg no HT	0.56	0.45	0.50
LogReg 1 HT	0.66	0.54	0.59
LogReg 2 HT	0.69	0.58	0.63
LogReg 3 HT	0.72	0.57	0.64
LogReg 4 HT	0.73	0.58	0.65
Lin. SVM no HT	0.58	0.31	0.40
Lin. SVM 1 HT	0.55	0.6	0.57
Lin. SVM 2 HT	0.64	0.58	0.61
Lin. SVM 3 HT	0.64	0.58	0.61
Lin. SVM 4 HT	0.69	0.61	0.65
RoBERTa no HT	0.67 ± 0.028	0.51 ± 0.027	0.58 ± 0.009
RoBERTa 1 HT	0.69 ± 0.024	0.56 ± 0.028	0.62 ± 0.019
RoBERTa 2 HT	0.72 ± 0.021	0.61 ± 0.034	0.66 ± 0.024
RoBERTa 3 HT	0.71 ± 0.015	0.66 ± 0.028	0.68 ± 0.010
RoBERTa 4 HT	0.74 ± 0.021	0.66 ± 0.038	0.69 ± 0.020
Twitter RoBERTa no HT	0.62 ± 0.027	0.51 ± 0.025	0.56 ± 0.007
Twitter RoBERTa 1 HT	0.68 ± 0.026	0.56 ± 0.009	0.61 ± 0.006
Twitter RoBERTa 2 HT	0.71 ± 0.014	0.57 ± 0.020	0.63 ± 0.016
Twitter RoBERTa 3 HT	0.71 ± 0.016	0.60 ± 0.022	0.65 ± 0.017
Twitter RoBERTa 4 HT	0.72 ± 0.026	0.63 ± 0.025	0.67 ± 0.008
RoBERTa only HT	0.49 ± 0.034	0.71 ± 0.021	0.58 ± 0.022

Table 1: Comparison of different classification methods. HT stands for Historical Tweets.

in tweets, both for transformer-based and traditional ML models. Importantly, in our study the benefits of incorporating user-context clearly outweigh the benefits of using domain-specific intermediate training (compare, e.g., 'RoBERTa no HT' vs 'RoBERTa 2 HT' – with statistically significant F1 improvement with p=0.004 – and 'RoBERTa no HT' vs 'Twitter RoBERTa no HT', resulting in a statistically insignificant decline in F1).

Interestingly, including the historical tweets levels the field between linear models and Transformers. Differences between either RoBERTa 4 HT and logistic regression 4 HT are not statistically significant for the RoBERTa models with median F1 (although the plain RoBERTa model with the highest F1 yields a 'statistically significant' improvement in an uncorrected test). We believe this can be explained by the transformer-based models being better at dealing with very sparse utterances of single tweets (both RoBERTa models with no HT are significantly more effective in terms of F1 than logistic regression without historical tweets; p=0.01 and p=0.05, respectively). The presence of additional contexts yields the dense representations used by RoBERTa to 'fill in the blanks' less useful.

Improvements in effectiveness comparable are in magnitude (although not directly comparable[2]) to improvements attained using more specialised models and user metadata on a super-set of the same data by Naseem et al. (2021). Harnessing topic-specific historical tweets can be seen as an alternative mechanism of user profiling, which arguably carries lower risk of re-identification than combining network feature, user metadata, and textual features.

The last row of Table1 reports an experiment with a model exposed to historical context only, both at training and testing. I.e., the task here can be represented as predicting the stance of the next tweet from a specific user, given their posting history on a specific topic (here, vaccines). Interestingly, it seems to be the only recall-biased model in our experiments, which indicates that the models are more likely to mistake vaccine-positive/neutral contexts for vaccine-negative contexts than they are to mistake a vaccine-positive-or-neutral tweet for a vaccine-negative one.

6 Limitations

The presented work constitutes an initial, exploratory step towards incorporating user-produced context (historic posts) into a vaccine stance surveillance pipeline. We only explore an artificial version of the problem, where we look at artificial user groups with 1 to 5 posts specific to the topic of interest.

Another limitation of our work is relates to limitations of current transformer-based classification models, which can only be applied to texts of limited length. Our exploratory study does not offer solutions towards incorporating broader historical context in training and inference.

7 Conclusions and Future Work

We investigated the problem of vaccination stance detection in Twitter using historical tweets by different Twitter users. We compared different text classification methods to identify stance of users. Our results point to a methodology to improve detection effectiveness through improved data collection pipeline for health-related social media *infoveillance*. We hypothesise that our strategy is especially applicable in scenarios where the public is highly polarised.

As future work, we will explore opportunities, and difficulties, around the use of user-generated context in experimental setup more similar to real-world applications.

[2]The authors of the cited work evaluated their methodology in a multi-class setup, and without filtering for historical tweet availability (thus, with more training data).

Acknowledgements

This work has CSIRO's ethics committee aproval (2021_115_LR). This work is supported by the CSIRO's Precision Health Future Science Platform.

References

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Douglas Biber and Edward Finegan. 1988. Adverbial stance types in english. *Discourse Processes*, 11(1):1–34.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Prateek Chaudhry and Matthew Lease. 2022. You are what you tweet: Profiling users by past tweets to improve hate speech detection. In *Information for a Better World: Shaping the Global Future*, pages 195–203. Springer International Publishing.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2021. Synthetic examples improve cross-target generalization: A study on stance detection on a Twitter corpus. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 181–187.

Mike Conway, Mengke Hu, and Wendy W. Chapman. 2019. Recent advances in using natural language processing to address public health research questions using social media and consumer generated data. 28(1):208–217.

Yingtong Dou, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2051–2055.

Adam G. Dunn, Didi Surian, Jason Dalmazzo, Dana Rezazadegan, Maryke Steffens, Amalie Dyda, Julie Leask, Enrico Coiera, Aditi Dey, and Kenneth D. Mandl. 2020. Limited role of bots in spreading vaccine-critical information among active Twitter users in the United States: 2017–2019. *American Journal of Public Health*, pages 319–325.

Pin Huang, Andrew MacKinlay, and Antonio Jimeno Yepes. 2016. Syndromic surveillance using generic medical entities on Twitter. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 35–44, Melbourne, Australia.

Adith Iyer, Aditya Joshi, Sarvnaz Karimi, Ross Sparks, and Cecile Paris. 2019. Figurative usage detection of symptom words to improve personal health mention detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1142–1147, Florence, Italy. Association for Computational Linguistics.

Antonio Jimeno Yepes, Andrew MacKinlay, and Bo Han. 2015. Investigating public health surveillance using Twitter. In *Proceedings of BioNLP 15*, pages 164–170, Beijing, China.

Aditya Joshi, Xiang Dai, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C Raina MacIntyre. 2018. Shot or not: Comparison of NLP approaches for vaccination behaviour detection. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 43–47, Brussels, Belgium. Association for Computational Linguistics.

Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C. Raina Macintyre. 2019. Survey of text-based epidemic intelligence: A computational linguistics perspective. 52(6).

Aditya Joshi, Ross Sparks, Sarvnaz Karimi, Sheng-Lun Jason Yan, Abrar Ahmad Chughtai, Cecile Paris, and C. Raina MacIntyre. 2020a. Automated monitoring of tweets for early detection of the 2014 Ebola epidemic. *PLOS One*.

Aditya Joshi, Ross Sparks, James McHugh, Sarvnaz Karimi, Cecile Paris, and C. Raina MacIntyre. 2020b. Harnessing tweets for early detection of an acute disease event. *Epidemiology*, 31(1):90–97.

Katherine Lanyi, Rhiannon Green, Dawn Craig, and Christopher Marshall. 2022. Covid-19 vaccine hesitancy: Analysing twitter to identify barriers to vaccination in a low uptake region of the uk. *Frontiers in digital health*, 3.

S. N. Medford, R. J.and Saleh, A. Sumarsono, T. M. Perl, and C. U. Lehmann. 2020. An Infodemic: Leveraging high-volume Twitter data to understand early public sentiment for the Coronavirus disease 2019 outbreak. *Open forum infectious diseases*, 7(7).

Tanushree Mitra, Scott Counts, and James Pennebaker. 2021. Understanding anti-vaccination attitudes in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 269–278.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology*, 17(3):1–23.

Roser Morante, Chantal van Son, Isa Maks, and Piek Vossen. 2020. Annotating perspectives on vaccination. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4964–4973, Marseille, France.

Usman Naseem, Matloob Khushi, Jinman Kim, and Adam G. Dunn. 2021. Classifying vaccine sentiment tweets by modelling domain-specific representation and commonsense knowledge into context-aware attentive gru.

Bahadorreza Ofoghi, Meghan Mann, and Karin Verspoor. 2016. Towards early discovery of salient health threats: a social media emotion classification technique. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 504–515.

Michael J Paul and Mark Dredze. 2017. Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 9(5):1–183.

Andrea Raballo, Michele Poletti, and Antonio Preti. 2022. Vaccine hesitancy, anti-vax, covid-conspirationism: From subcultural convergence to public health and bioethical problems. *Frontiers in Public Health*, 10.

Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2017. Automatic detection of stance towards vaccination in online discussion forums. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media*, pages 1–8, Taipei, Taiwan.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 778–783.

Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. 2018. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *European Conference on Computer Vision*, pages 87–104.

Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In *The 26th International Conference on Computational Linguistics*, page 3485–3495.