Detecting Vaccine Skepticism on Twitter Using Heterogeneous Information Networks

Tim Kreutz[1[0000−0001−9633−1995]] and Walter Daelemans[1[0000−0002−9832−7890]]

CLiPS - Computational Linguistics Group, University of Antwerp, Antwerp, Belgium
tim.kreutz,walter.daelemans@uantwerpen.be
https://www.uantwerpen.be/en/research-groups/clips/

Abstract. Identifying social media users who are skeptical of the COVID-19 vaccine is an important step in understanding and refuting negative stance taking on vaccines. While previous work on Twitter data places individual messages or whole communities as their focus, this paper aims to detect stance at the user level. We develop a system that classifies Dutch Twitter users, incorporating not only the texts that users produce, but also their actions in the form of following and retweeting. These heterogeneous data are modelled in a graph structure. Graph Convolutional Networks are trained to learn whether user nodes belong to the skeptical or non-skeptical group. Results show that all types of information are used by the model, and that especially user biographies, follows and retweets improve the predictions. On a test set of unseen users, performance declines somewhat, which is expected considering these users tweeted less and had fewer connections in the graph on average. To consider multiple degrees of vaccine skepticism, the test set was annotated with more fine-grained labels and the model was repurposed to do multiclass classification. While the model trained on binary labels was unsuited for this additional task, heterogeneous information networks were found useful to both accurately model and visualize complex user behaviors.

Keywords: Vaccination · Stance detection · Social media · Twitter · Graph Convolutional Networks

1 Introduction

The vaccine roll out during the COVID-19 pandemic has posed numerous diplomatic and societal challenges for governments around the world. Now that vaccines are available in most countries, significant parts of the world population remain unvaccinated not for logistical reasons but because of vaccine skepticism.

Vaccine skepticism is defined as a delayed acceptance of the COVID-19 vaccine, or the outright refusal to take a vaccine despite its availability [7]. Willingness to get vaccinated has become a polarizing issue and its contention has increasingly moved to online spaces. Social media in particular, are used to share not only personal opinions, but information from family, friends, health professionals, local governments and a mixture of accurate and questionable news outlets.

A 2020 study found that vaccine hesitancy is more prevalent amongst people who list the internet as one of their main sources of medical information [4]. The proliferation of questionable information, ranging from misleading news to conspiracy theories about the ongoing pandemic, has lead to a social media *infodemic* [7]. Especially in times when skepticism of government measures can cause difficulties in overcoming a crisis, it is important to identify and understand the users who believe such misinformation.

This paper sets out to identify Dutch speaking Twitter users who are skeptical of vaccination against COVID-19. We choose Twitter as a platform because its user profiles as well as most user interactions are publicly available. We target Dutch speaking users because the group of anti vaccination users is still marginal compared to the United States [11], making it feasible to scale our tool to give a complete picture of anti vaccination groups in the two countries.

Our aim is to reflect the dynamic nature of the vaccine skeptics both in their behaviors and communications by acknowledging that degrees of skepticism exist. We therefore implement a heterogeneous information network which models a user's actions as well as their language. Users actively decide who to follow or retweet, and tweet out in support or opposition of ideas. All of this information is relevant to positioning users relative to each other and identifying their group membership.

1.1 Contributions of this work

The aim of this work is to identify users who are skeptical of vaccines on Dutch speaking Twitter. We specifically answer the following three questions:

1. Which behaviors convey vaccine skepticism on Twitter?
2. Can we improve detection of skeptical users by using a combination of linguistic and network features in a heterogeneous information network?
3. Can a model trained for making the binary distinction between skeptical and non-skeptical users be reused to identify degrees of skepticism?

Our contribution is to develop an accurate model which takes into account network features and linguistic cues for classifying Twitter users as skeptical of vaccines or not. The model output warrants further research into the spectrum of skepticism, which can range from hesitant users who simply ask questions about the vaccines to anti-vaccination users who actively spread conspiracy theories amongst their followers.

The next section will discuss related research to contextualize this work. We explain how we constructed our corpus and how it is fit into the mold of a heterogeneous information network in the methodology section. We first develop a model to make binary distinctions between users who are skeptical of vaccination and those who are not. The model output is evaluated in terms of its discriminative accuracy and performance outside the initial dataset on unseen users. We then discuss our findings with an error analysis and consider how the binary model could be used to target more specific groups of users on the spectrum of vaccine skepticism.

2 Background

Detecting stance on Twitter has been frequently organized as a shared task on a wide range of polarizing topics, including vaccination. The last iteration of the stance detection tasks at IberLEF [1] provides meta-information of messages and users to incorporate social features in the submitted systems. These features could however not be used to connect users and messages in a network of data to exploit graph edges in that way.

This work fits a larger context of research done to improve understanding of negative opinions about vaccines online. The Vaccine Confidence project (VCP) [1] for instance, has monitored concerns over vaccines in media reports since 2012 to show where and when specific concerns arise. Most notably, it tracked opinions about vaccination for the influenza A(H1N1) virus outbreak in 2009, for the human papillomavirus (HPV) and for coronavirus SARS-CoV-2.

Previous work on Twitter data has cast detection of vaccine skepticism as a text categorization problem. [6] collected 6,000 tweets about HPV to better understand the low vaccination coverage in the U.S. Their setup used SVMs with basic n-gram features to determine whether a tweet was positive, negative or neutral about vaccination for HPV.

[9] translated this approach for the context of COVID-19 in the Netherlands with the goal of identifying tweets with a negative stance towards vaccination. The study finds that identifying such tweets is a non-trivial problem, due to the many motivations for adopting a negative stance and the relative scarcity of the negative class, especially when compared to a larger community of anti vaccination users in the U.S. context.

Vaccine skepticism can alternatively be cast as a social phenomenon which exhibits itself in complex user interactions, rather than in the broadcasting of a static opinion. [13] focus more on network dynamics when tracking sentiments about the H1N1 vaccine on Twitter. Negative sentiment is observed to be more contagious than positive sentiment, but a larger opinionated neighborhood inhibits this contagion. The results in [13] suggest that vaccine skeptic content is mostly circulated in small groups of homophilic Twitter users.

This is further underlined by a recent report by the Center for Countering Digital Hate (CCDH) entitled The Disinformation Dozen [5]. In the report a massive stream of anti vaccination misinformation on Facebook and Twitter is shown to originate from only twelve influencers.

Beyond just describing network features, [2] aim to discover such clusters on Twitter with several existing community detection algorithms. They find negative sentiments to persist in groups of few users that are not well connected. In a preliminary geographical analysis, these negative networks seem to operate mostly in U.S. territories. Although the setup is useful for identifying several key drivers of vaccine skepticism on Twitter, it misses those users that are simply hesitant about vaccination.

[1] https://www.vaccineconfidence.org/

We primarily differ from previous setups by framing vaccine skepticism as a social problem at the level of the user. Although identifying skepticism on a document level can help detect skeptical users, it ignores other public user information. Especially on Twitter, where messages are short and lacking of context, using all available information albeit tweets, public user profiles or interactions with others, is important to develop an accurate model. Conversely, algorithms classifying subgraphs, such as in community detection, do identify drivers of vaccine skepticism using network features, but will be less useful for finding hesitant users, who would more likely operate in the periphery of such communities.

The current work is situated at the interface of text categorization and network analysis by employing recent graph based methods which work both with content as well as network features. Methodologically, our work is closest to [3] who classify a tweet on the basis of its content, the content of previous tweets by the same user, and a user representation based on the reply network of the tweet. Adding the network representation lead to substantial improvements in the accuracy of a logistic regression classification model.

3 Methods

Twitter allows diverse interactions between users, all giving different signals about the social relationships between them. Our methodology is derived from this social nature of the data, modelling the interactions between users as faithfully and completely as possible.

Network graphs are suitable for modelling many real-world data, social networks among them [14]. Starting from the straightforward and homogeneous network of followers, we increase complexity of the graph representation by introducing encoded user biographies and tweets. As such, we end up with a heterogeneous information network which inherently stores network features and contains linguistic features at the node level; User behavior is modelled as actions ("*who do users follow?*") and words ("*what do users write?*").

We choose Graph Convolutional Networks (GCNs) to do node prediction for the unlabeled samples. GCNs are a Convolutional Neural Networks (CNN) implementation operating directly on graph-structured data [8]. GCNs learn a function of signals and features on a graph by optimizing cross-entropy loss on labelled samples. An unlabelled node is passed signals from nearby nodes, using the learnt weights to predict its class.

The GCN algorithm scales linearly for graphs with huge numbers of nodes and edges, and accurately learns node representations from both graph structure and neighboring node features. With only a very limited number of nodes annotated ($<$10%), GCNs significantly outperformed other graph-based algorithms on a range of graph-learning tasks, node prediction among them [8].

Relational Graph Convolution Networks (RGCNs) [14] are a necessary extension of the GCN approach to operate on graphs with multiple types of nodes and edges, or heterogeneous graphs. Node representations are created by merging signals from different edge types. To return to the case of a social network: an

unlabeled user node is passed the features from its own authored tweets, tweets it has retweeted, and from users in its social network.

We describe the corpus of Twitter data below. The experiments were designed to first learn prototypical distinctions between users who are skeptical of COVID-19 vaccinations and others who frequently tweet about vaccination. We then expand the data to show how our prototypical model can be applied to unseen users. In an additional task, we aim to repurpose the outputs of the binary model to identify multiple degrees of skepticism. As such, it could identify users who position themselves on the edge of radicalization, being themselves hesitant of COVID-19 vaccination and in search of answers.

3.1 Data

The training data is taken from a snapshot of Dutch tweets from January 2020 to June 2021. A filter was applied based on a regex that was designed to retrieve mentions of vaccines and relevant pharmaceutical companies. In the chosen period, we found 660,415 users having at one point tweeted about these topics. To make annotating an initial set of users feasible, we further limited the group to those users who posted an original tweet about the relevant topics at least once in each of the 18 months included in the dataset; The 3,565 users that were found this way can easily be inspected by hand to assign accurate labels, and are guaranteed to actively have taken a stance on COVID-19 vaccination.

Annotation task An annotation task was set up to assign labels to the 3,565 users. A single annotator was shown the biographic text from a user's Twitter profile as well as three randomly chosen tweets about vaccines from their timeline. If the annotator saw clear evidence of vaccine skepticism either in the biography or in the written tweets a `SKEP` label was assigned. If the biography and initial three tweets did not give clear indication of a stance either way, three more tweets could be shown. Eventually, if no clear signs of skepticism were found in any of the tweets, the user is assigned to the non-skeptical group by default.

The result is a dataset with 1,781 skeptical users, and 1,784 non-skeptical users. We first extract a list of followers for each user, and model a homogeneous graph with users as nodes and follows as edges. An extract from the network of training data is shown in Figure 1a.

Biographies We then implemented encoded user biographies as node features. Most users (84%) in the annotated data added a biography to their profile. The short texts were encoded using Sentence-BERT [12], a state-of-the-art language model for sentence encoding. The specific model (all-mpnet-base-v2) was picked for its accuracy in diverse use cases, as well as its relative speed which becomes important when encoding a large number of tweets in later steps.

Tweets The final graph representations implement tweet nodes. Users connect to tweet in one of two ways: either by having written a tweet, or by having retweeted a tweet. Tweet nodes carry their Sentence-BERT encodings as feature values. Convolutions by the GCN over edges are performed separately for each edge type, making the graph in Figure 1b truly heterogeneous. We follow the same iterative increasing of complexity in the graph in the results section to see how the heterogeneous information contributes to user node classification.

Training data		Test data	
Skeptical	1,781	Anti	627
		Hesitant	297
Other	1,784	Unknown	220
		Pro	404
Total	3,565	Total	1,548

(a) Labels

	Training data	Test data
Users	3,565	1,548
Tweets	530,935	111,177
Follows	353,074	34,384
Writes	530,935	111,177
Retweets	303,523	23,338

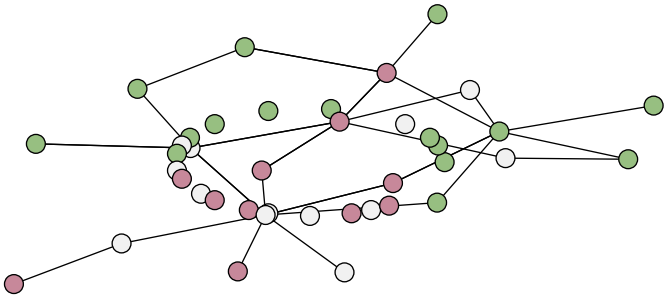
(b) Network information

Table 1: Differences between the data used to train a prototypical model of learning vaccine skepticism and the test data consisted of unseen users. The test users are generally less active (writes and retweet edges) and less connected (follow edges). More fine-grained labels were assigned to the test group to evaluate the suitability of the prototypical model for detecting degrees of skepticism.

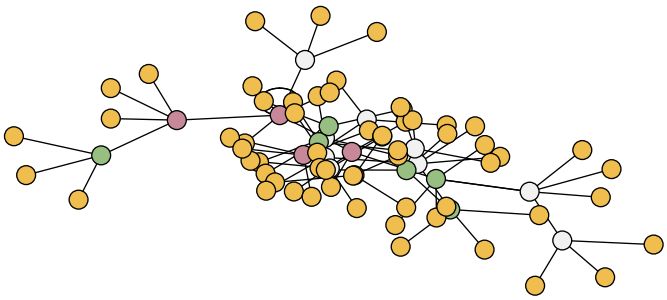
Unseen users The pre-filtering that was applied when selecting users for training may restrict how well the model generalizes to unseen users. We selected 1,548 Twitter users randomly, placing only the restriction that they have at one point tweeted about vaccination, as test data to evaluate whether the model will generalize to users that may show less evidence of skepticism simply because they are less active and well-connected in the network.

The test set received more fine-grained labels to better reflect different degrees of skepticism. We took definitions from [10] and divided vaccine skepticism into two classes: anti vaccination and hesitant. The non-skeptical users were divided into pro vaccination users and unknown users. The differences between the data used for training the model and the test data is described in Table 1.

To translate to a multiclass setting, we take the softmax output of the binary model and subtract the probability for the positive class (SKEP) from the negative. The resulting predictions form a distribution ranging from -1 to 1. The model, being unaware of prior class distribution over unseen users, heuristically creates buckets of samples per quantile, corresponding to their respective labels: ANTI, HES, UNK, and PRO.



(a) Extract from the homogeneous graph with user nodes and following as edges. The complete graph consists of 5,113 user nodes with 387,557 follow edges between them.



(b) Extract from the heterogeneous graph with user and tweet nodes. The complete graph consists of 5,113 users, 642,112 tweets, 387,557 follow edges, 642,112 write edges and 326,861 retweet edges.

Fig. 1: Steps of increasing complexity in a network modelling Twitter users central to discussions about vaccines.

4 Results

The Graph Convolutional Networks were trained with default parameters from the implementation in [8]: one hidden layer, .01 learning rate, 50% dropout for 200 epochs. Signals passed from different relation types in the heterogeneous setting were grouped by summing, then reduced by taking the mean of the different signals like the standard implementation in [14].

4.1 Graph performance

Increasing the graph complexity does not necessarily instill more useful information in the network. Tweets about topics other than COVID-19 vaccinations may introduce noise while retweeting other users may not always signal support for their message. Table 2 shows the performance of individual graph implementations on a validation set. We vary the percentage of supervised nodes that propagate information elsewhere in the network to see the effect of training size. The size of the validation set inversely consists of the remaining unsupervised user nodes.

The first graph implementation is featureless (see Figure 1a), meaning GCNs rely only on network information to induce node embeddings. Much improvement can be made by using encoded biographies as user features (#2). This result is especially impressive considering nearly 16% of users in the training data and validation data does not have a biography in their profile.

The network structure in graph #3 is further enriched by incorporating retweets. Users that did not previously share any formal ties, can convey similarity by retweeting the same tweet. Finally, encoded tweet features and write edges to propagate them are implemented in graph #4 (Figure 1b).

Increasing the training size had a positive effect for the graph implementations with more (types of) edges and nodes. This is somewhat expected as diversity of examples increases with network complexity. However, the effect was surprisingly small. GCNs that trained only on 20% of the available annotations (713 users), performed nearly as well as those trained in four times as much training samples. In fact, the featureless implementation performed better when using only few samples.

4.2 Test results

The results on test users drop off somewhat since the method by which they were selected differs from the users in the training and validation data. Whereas the latter group tweeted about vaccination more actively and is well-connected in the network, as partly reflected in the network information in Table 1b, the test setting simulates plugging any unseen Twitter profile into the network and asking our classifier about their stance on vaccination. Not only could the GCNs have fewer information to work with, the very label may be more doubtful as the user may not take a firm stance one way or the other.

Graph	Features	Edges	20%	40%	60%	80%
#1	-	F	**.647**	.624	.643	.615
#2	U	F	.803	.793	.793	**.808**
#3	U	F + R	.861	.860	.868	**.871**
#4	U + T	F + W + R	.862	.871	.874	**.879**

U = User features, T = Tweet features,
F = Follow edges, R = Retweet edges, W = Write edges

Table 2: Node prediction accuracy on validation data. Improvements were made by iteratively increasing the complexity of the graph representation, first adding user features (biographies), retweet relationships, tweet features and write edges. There was a limited effect of training size, affirming that GCNs work well even with very limited supervision.

Still, the graph implementation that used both user and tweets features, follow, write and retweet edges, trained on all available training data, was able to assign a correct binary label to unseen users in **74.1 percent** of cases.

Degrees of skepticism The confusion matrix in Figure 2 shows the types of errors made by the classifier in the binary setting, resulting in the accuracy mentioned above, as well as in the setting with a multi class distinction. As expected, anti vaccination and pro vaccination sentiments, representing the extremes of the scale, are easier to predict, while vaccine hesitancy is predicted correctly as skeptical in only 62 percent of cases.

	ANTI	**HES**	**UNK**	**PRO**	Total
SKEP	448	185	86	24	733
OTHER	179	112	134	380	815
Total	637	297	220	404	1,548

Fig. 2: Confusion matrix comparing binary output from the first model to the true fine-grained labels, representing degrees of skepticism in the test data.

Note again that the model is not trained to do these multiclass predictions, as annotating sufficient data for training with such subtle distinctions becomes prohibitively time consuming. However, model confidence may serve as a proxy for identifying vaccine hesitancy, a heuristic based on how annotators themselves have trouble assigning labels to this group.

Table 3 shows the results of the softmax discretization heuristic. Again, the ANTI and PRO class are more accurately predicted than the HES and UNK users. We hypothesize that the binary model could recall more HES users, giving it purpose as a prefiltering model for further annotation or silver-labelling of a dataset. However, this did not turn out to be the case. Further development would be needed to employ the GCNs in this multiclass setting.

Class	P	R	F
ANTI	0.643	0.397	0.491
HES	0.222	0.290	0.251
UNK	0.142	0.250	0.181
PRO	0.509	0.488	0.499
AVG	0.388	0.373	0.368

Table 3: Detailed results for each class in the multiclass setup. Discretizing softmax to multiple degrees of vaccine skepticism did not yield good results for the more subtle groups of hesitant users, and users who were labelled as unknown.

5 Conclusion

In this paper we explored the ways in which Twitter users may show vaccine skepticism: through linking up with other users, by spreading their message, by communicating their stance in a self authored biographic text or by tweeting about vaccines.

Our main contributions stem from approaching the task of detecting vaccine skepticism as a problem at the user level. Abstracting away from individual documents allows connecting diverse types of information and outperforming the models who draw information only from text. Heterogenous graph representations are ideal for modelling both social action and linguistic features, and a new vein of neural graph-based models learn weights accurately and rapidly on these graphs, even with millions of nodes.

This heterogeneous information was modelled on a dataset of Dutch speaking Twitter users who tweet relatively often about vaccines. We trained Graph Convolutional Networks for predicting node labels, incrementally feeding it more types of edges, nodes and features.

The most complex graph, which incorporates user and tweet nodes and connects them through follow, write and retweet edges performed best on the validation data. The effect of training size was limited, showing that GCNs trained on few training samples are able to generalize well.

Each edge thus contributed to a better model. Biographies notably contained a lot of information directly and indirectly indicative of skepticism, shown in a huge leap in performance. Linking users to tweet they had retweeted was also

very beneficial to the model, while write edges yielded limited improvement since they contributed only one edge.

In the test setting we implanted unseen users in the network that have at one point tweeted about vaccination in our data set. Results for this group were lower, although still impressive at 74.1% accuracy considering these nodes provided less context.

Unfortunately, the prototypical model could not be repurposed for detecting degrees of skepticism. The GCNs showed more confidence in predicting anti vaccination and pro vaccination users, but using these confidence scores to identify the group of hesitant users proved non-trivial. Future research could therefore focus on tuning a model specifically towards identifying hesitancy in vaccine stance.

Developing graph-based deep learning models is another useful avenue for future work. Especially in the case of complex user behaviors, which are often discretized on social media, GCNs show they can exploit network and node features in a complementary way. To encourage new work and the reproducibility of the current work, our code and the data representations are available from our repositories[2] upon request.

References

1. Agerri, R., Centeno, R., Espinosa, M., de Landa, J.F., Rodrigo, A.: Vaxxstance@ iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. Procesamiento del Lenguaje Natural **67**, 173–181 (2021)
2. Bello-Orgaz, G., Hernandez-Castro, J., Camacho, D.: Detecting discussion communities on vaccination in twitter. Future Generation Computer Systems **66**, 125–136 (2017)
3. Béres, F., Csoma, R., Michaletzky, T.V., Benczúr, A.A.: Vaccine skepticism detection by network embedding. arXiv preprint arXiv:2110.13619 (2021)
4. Charron, J., Gautier, A., Jestin, C.: Influence of information sources on vaccine hesitancy and practices. Médecine et Maladies Infectieuses **50**(8), 727–733 (2020). https://doi.org/https://doi.org/10.1016/j.medmal.2020.01.010, https://www.sciencedirect.com/science/article/pii/S0399077X20300457
5. for Countering Digital Hate, C.: The disinformation dozen: Why platforms must act on twelve leading online anti-vaxxers. Counterhate.com (2021)
6. Du, J., Xu, J., Song, H., Liu, X., Tao, C.: Optimization on machine learning based approaches for sentiment analysis on hpv vaccines related tweets. Journal of biomedical semantics **8**(1), 1–7 (2017)
7. Hughes, B., Miller-Idriss, C., Piltch-Loeb, R., White, K., Creizis, M., Cain, C., Savoia, E.: Development of a codebook of online anti-vaccination rhetoric to manage covid-19 vaccine misinformation. medRxiv (2021). https://doi.org/10.1101/2021.03.23.21253727, https://www.medrxiv.org/content/early/2021/03/26/2021.03.23.21253727
8. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)

[2] https://github.com/clips/vaccine_skepticism

9. Kunneman, F., Lambooij, M., Wong, A., Van Den Bosch, A., Mollema, L.: Monitoring stance towards vaccination in twitter messages. BMC medical informatics and decision making **20**(1), 1–14 (2020)
10. Larson, H.J., Broniatowski, D.A.: Volatility of vaccine confidence. Science **371**(6536), 1289–1289 (2021). https://doi.org/10.1126/science.abi6488, https://science.sciencemag.org/content/371/6536/1289
11. Mitra, T., Counts, S., Pennebaker, J.W.: Understanding anti-vaccination attitudes in social media. In: International Conference on Web and Social Media (ICWSM). AAAI, AAAI (May 2016), https://www.microsoft.com/en-us/research/publication/understanding-anti-vaccination-attitudes-in-social-media/
12. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2020), https://arxiv.org/abs/2004.09813
13. Salathé, M., Vu, D.Q., Khandelwal, S., Hunter, D.R.: The dynamics of health behavior sentiments on a large online social network. EPJ Data Science **2**(1), 1–12 (2013)
14. Schlichtkrull, M., Kipf, T.N., Bloem, P., Van Den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: European semantic web conference. pp. 593–607. Springer (2018)