scientific reports

Check for updates

OPEN

Leveraging complex network features improves vaccine stance classification

Durazzi Francesco[1,4], Barbieri Niccolò[1,4], Faccin Mauro[1,2,4], Gori Davide[3] & Remondini Daniel[1]✉

The widespread use of social media allows unprecedented ways to monitor opinions and stances regarding critical public health issues globally. Advanced Natural Language processing algorithms are being used routinely to extract information and classify vaccination hesitancy or stance. However, communication on online social networks such as Twitter (now X) is carried by short messages, the meaning of which can be difficult to understand in the absence of context. Therefore, in this study we propose the use of complex-network features extracted from the social network to integrate and enhance text-based Deep Learning models. Leveraging a dataset of about 20 million Italian language posts (of which about 7000 were manually annotated), we showed how the integration of text and network features improves vaccine stance classification, especially for the most polarized classes. Additionally, network features overperformed text features in a dataset collected a year after model training, possibly indicating how the social network changes more slowly than the trending words or topics.

With more than 7 million deaths reported by WHO as of March 2025, COVID-19 is marked as one of the most severe pandemics in modern history. The velocity of the epidemic spread and the rapid emergence of severe respiratory symptoms highlighted the need for worldwide impactful and timely vaccination strategies to increase immunization in the population or at least decrease the severity of the illness[1]. The delivery of scientific information displayed unprecedented mechanisms, sometimes with contrasting messages, such as promoting the usage of hydroxychloroquine without proven benefits[2]. Nevertheless, this helped to raise awareness and enforce infection control measures, but also added pressure for proper interpretation by the general public[3], in particular in demographic contexts where health literacy is low, as in Italy[4]. Vaccines have been developed rapidly and approved through the Emergency Use Authorization[5] but, despite the enforcement of vaccinations in many countries, vaccine hesitancy and opposition posed a significant health risk at individual and population level[6]. The willingness to accept COVID-19 vaccines is related to many cognitive, psychological, socio-demographic and cultural factors, and the assessment of the diffusion in the population of negative stance towards vaccination is crucial to guide interventional measures aimed at ensuring high-level of public health[6–8].

Social media do not only offer access to monitor vaccine hesitancy worldwide[9,10] but also may play an active role in propagating (mis)information and polarizing views. For COVID-19 in particular, an increasing opposition to vaccines has been observed on Twitter during the first two years of the pandemic[11]. Social network discussions and opinions do not live in a parallel world apart from the offline one; in fact, social media monitoring during the pandemic has already shown many correlations of sentiments towards vaccination with country-specific events related to the vaccination programme[12]. While it is argued how much the different online social networks may be representative of the population, a vast majority of internet users search health information online[13,14], thus increasing the risk of exposure to biased (mis)information. A relevant trend in the vaccine discourse on Twitter is the prevalence of negative sentiment also in pro-vax communities of users, as it has been observed already in 2020[15]. According to ECDC, the online spread of rumours surrounding vaccination, including adverse events following vaccination, has contributed to the growth of vaccine hesitancy and in some cases may have contributed to disease outbreaks in unvaccinated populations[16].

Many attempts have been done to support public health agencies in monitoring vaccine stance over social media, aimed to automatizing the classification of social networks posts and users to estimate prevalence of AntiVax views[17–20], identify echo-chambers and bubbles spreading misinformation or uncertainty[21,22], and

[1]Department of Physics and Astronomy, University of Bologna, Bologna 40127, Italy. [2]Department of Physics and Astronomy, University of Padova, Padova 35122, Italy. [3]Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna 40127, Italy. [4]These authors contributed equally: Durazzi Francesco, Barbieri Niccolò and Faccin Mauro. ✉email: daniel.remondini@unibo.it

possibly gain knowledge about weaknesses of the vaccination communication strategies. Successful vaccination programmes are built on understanding and taking into account individuals' and communities' beliefs, concerns and expectations regarding the vaccine and the disease[23]. Given the observation that Pro and AntiVax social media posts span over different topics and display clear patterns marking them recognizable[24], many studies opted for machine learning classifiers grounded in Natural Language Processing (NLP) to extract information from textual data and classify the posts based on their stance[25–27]. Since their introduction, deep learning Language Models (LMs) based on the attention mechanism, named transformers[28], are leading the race for automatized text classification[29–31]. Some studies claim almost perfect accuracies in detecting vaccine stance on english-language Twitter posts[18,19] although preprocessing and filtering steps can limit the generalizability of the performance to real-life scenarios: tweets with irony or sarcasm may have been manually removed[18] and broadness and complexity of the views may have been reduced by limiting the tweets to specific subtopics (e.g. maternal vaccinations[19].

It can indeed be very difficult to resolve the stance of certain tweets in the absence of context[19,32], so any attempt to circumstantiate and contextualize a social media post can help. In this study, we propose to address the issue by providing information about the users inside the social network itself. Complex network analysis has been applied many times in the context of social media monitoring for vaccine hesitancy, showing that communities of tightly-connected users may be related to vaccine stance[22], and that vaccine hesitancy is associated with pre-existing political or religious beliefs[33]. The interplay between social network structure and vaccine stance is bidirectional, because on the one side the expression of negative opinions on vaccination is associated with previous exposure to negative stance within the network[14], but on the other side also strong stance positions influence and reshape the network density[34].

For this reason, we hypothesized that integrating the text of tweets with contextual information encoded in the network structure may increase the ability to classify vaccine stance. To test our hypothesis, we employed a dataset of over 19 million tweets in Italian language related to vaccination, collected along the COVID-19 pandemic, partly annotated for a previous study (approximately 7 K tweets)[32], partly annotated for this study (approximately 700 tweets), and the remaining utilized to build the network structure and for domain-specific re-training of the text language model. By comparing vaccine stance classifiers for tweets using as inputs text and network features (separately or together), we observed that employing network features for vaccine stance classification (1) increased classification accuracy when integrated to text-only classifiers, (2) overperformed text classifiers in tweets collected more than a year after training and (3) performed optimally in presence of highly-polarized discussion (i.e. with high prevalence of non-neutral tweets).

Methods

Data

We gathered almost 20 million tweets using Tweepy, a specific Python package, from October 2020 during the pandemic's peak until March 2023 (Supplementary Figure S1). We used the continuous stream filter endpoint of the Twitter API with the academic license, which was available at that time. The tweets were chosen from Italian Twitter using specific search words, namely: "vaccino", "vaccini", "vaccinazione", and "vaccinazioni", which are the Italian translations of "vaccine(s)" and "vaccination(s)". A subset of tweets was randomly sampled for manual annotation of vaccine stance, while the full set of 20 M tweets hasbeen used to build the retweet network (see Network features and classifier in the Methods section). The first annotation process involved 6508 randomly selected tweets, up to January 2021 ("first dataset" from now on). A second annotation was performed on 715 tweets from a later period, between July 2022 and March 2023 ("second dataset" from now on). The tweets from the second dataset were used to test the model's robustness against time drifts. The annotation process of the dataset by three separate manual annotators for each tweet has already been detailed in a previous study[32] and aimed to assign a vaccine stance label to each tweet based on the text written by the user. Although it is reasonable to assume that most tweets about vaccination in the collection period referred to COVID-19, the relevance of all tweets to the COVID-19 theme could not be ensured, so annotators could only assess the stance against a generic vaccine. Each of the annotated tweets has a unique label, being: "ProVax" if the tweet encouraged the adoption of the vaccine or reported personal stories of having been vaccinated; "AntiVax" if the tweet expressed skepticism or aversion to the use of vaccines; "Neutral" if otherwise, including many news reports and mentions.

Of the first dataset 6508 tweets were annotated, 1570 of these were labelled as AntiVax, 952 as ProVax, and 3986 as Neutral. In the second dataset, 715 tweets have been annotated exclusively for this study with the same procedure as the first dataset; 36 tweets were labelled as ProVax, 170 as Neutral and 509 as AntiVax.

Due to the highly imbalanced distribution of classes, with neutral and anti-vaccine annotations being overrepresented, random undersampling was utilized to create a balanced dataset for model training. We then split the first dataset into a training set and a test set, randomly. We made use of 5-fold cross-validation (CV) on the training set to determine the best-performing classifiers. We then re-trained each selected classifier on the full training set before evaluating the performances on the test set and on the second dataset, both completely independent from the training set. Bootstrapping 10,000 times was used to estimate a confidence interval of the classifiers' performance on the test set and on the second dataset.

Network features and classifier

We built a retweet network at user level, where nodes represent users while a link from user i to user j represents the number of times user j retweeted a tweet written by user i[35]. After cleanup and extracting the giant component, the two datasets are composed of 1.40 M and 13.34 M retweeted tweets from 149k to 456k users respectively. We used network measures that extract information from the network topology and quantify the proximity between users. It is important to note that network analysis focuses on users rather than individual tweets, meaning that each tweet from the same user shares the same network information, thus having identical

network features utilized for tweet classification. Over time, the network grows by cumulating new retweets, which means creating new nodes and links or increasing link weights.

Community participation ratio

Users of online social networks often self organize in tightly connected groups or communities with weaker connection to other parts of the network. This mesoscopic structure highlights groups of similar individuals. For each user, we determined the strength of community membership to the largest communities. To extract the community structure of the network we used the Leiden algorithm and the Label Propagation algorithm from the Python module igraph. The former proposes a greedy approach to modularity maximization and was introduced as an improvement over the classical Louvain algorithm[36]. The latter labels each node by identifying the dominant community in its neighborhood[37]. We limited our analysis to the largest communities such that they jointly cover at least 90% of the network. From the community structure provided by the two algorithms, for each user *i* we extracted the fraction of neighbours $n_{i,\alpha}$ assigned to community α. We named these user-level features *norm_leiden* and *norm_lab_prop*, and refer to them as *community participation ratio*. When expanding the network with the second dataset, we maintained the same community structure of the initial dataset. In the second dataset network, for each user we computed the fraction of its neighbours assigned to the communities computed on the first dataset, hence only edges toward users already present in the first formulation were counted.

ForceAtlas2 (fa2)

ForceAtlas2 algorithm is a network layout algorithm which builds two-dimensional coordinatesfor the nodes on a force-based approach[38], optimized for fast computation in Python (https://github.com/AminAlam/fa2_modified). Specifically, it iteratively simulates attractive spring-like and repulsive gravitational-like forces between nodes, producing spatial embeddings where highly connected nodes are placed closer together[38]. To optimize the position of users in the second dataset, we executed the algorithm initializing the values to the original position, if available, or random, for new users. We named this two-dimensional feature set *fa2*.

Laplacian eigenvectors

We use the absolute values of the 10 smaller eigenvectors of the Normalized Laplacian (discarding the smallest $\lambda_0 = 0$). These spectral embeddings capture the network's mesoscopic structure — such as community organization and smooth variations in connectivity, which are informative for downstream analysis[39]. The eigenvectors have been computed on Octave from the symmetrized graph adjacency matrix. We named this feature set *norm_lap*.

Node2vec

Node2vec is an algorithm that generates low-dimensional embeddings for nodes in a network using a biased random walk approach[40]. In this process, each node is treated as a word, and the random walk is interpreted as a sentence, with embeddings created through a skip-gram method. Network embedding for the two datasets is computed independently, using the Python implementation in https://github.com/eliorc/node2vec. For this implementation, we have set the dimension of the desired embedding to 10. We named this feature set *n2v*.

Network-based classifier

The network features above were computed at user-level and were used as predictors for tweet vaccine stance. We employed logistic regression, ridge classification, and a dense neural network as classifiers, all implemented with 5-fold cross-validation. Within each validation fold, we centered the features around the fold training set center of mass, with the StandardScaler function of scikit-learn.

Text classifier

For text classification, we tested two Deep Learning language models, namely Italian BERT base (https://github.com/dbmdz/berts) and AlBERTo[26], both implemented through the HuggingFace transformers library in Python. BERT is a bidirectional Transformer-based model pretrained with a masked language modeling objective, enabling it to learn deep contextual representations of words from large unlabeled corpora[28]. Both these models were pre-trained on large corpora of Italian text: the first one was pre-trained on a large dump of the Italian version of Wikipedia, while AlBERTo was specifically pre-trained on italian tweets. Both these models were successively fine-tuned by us on our first dataset of 1.4 M tweets collected up to January 2021, performing a domain adaptation, to improve the performance on the specific field of COVID19 tweets. To do this, we identified the most common words from our tweet datasets that were absent from BERT's vocabulary, added them to the vocabulary, and then trained the classification model. To determine the best model between the BERT and AlBERTo, we compared their vaccine stance classification accuracy a 5-fold cross-validation on the training set.

Integrated classifier and performance evaluation

We selected the best network classifier and the best text classifier based on the K-Fold cross-validation as detailed in the Data paragraph of the Methods section.

We have compared these models' performance using accuracy, Matthews coefficient, per-class F1 scores and average F1 score, intended as the mean of the per-class F1 scores. The actual choices of models and hyperparameters operated on the training set through CV were operated to maximize the CV accuracy because the training set is perfectly balanced.

We then built an integrated classifier with a mixture-of-experts approach[41], combining the predicted class probabilities of the network and text classifiers. The two sets of predicted probabilities were summed with a

relative weight *w*. A weighted average of probabilities estimated by two different classifiers is the simplest and most interpretable way of combining them. The optimal value of *w* has been chosen as the one maximizing the accuracy on a 5-fold CV in the 1st dataset training set.

To compare the performance of the network classifier, the text classifier, and the integrated classifier, we computed the predictions on the 1st dataset test set and evaluated the aforementioned metrics.

To test the resilience of the classifiers across the passing of time and eventual concept drift, we repeated the evaluation on the 2nd dataset, containing tweets collected and annotated more than a year later than the 1st, as detailed in the Data paragraph of the Methods section.

Results

We developed various network features, beginning with the clustering of users into communities, specifically utilizing the Leiden algorithm. We decided to retain communities that together account for at least 90% coverage of all the users, which resulted to be six communities plus one for the excluded elements. The distribution is illustrated in Fig. 1.

In Fig. 2, it's shown the network layout computed with fa2 and colored according to community membership. The distribution of fa2 user coordinates displays two main blocks separated over the y-axis, with one of them mainly composed of one community.

Once we built a set of network features and text embeddings, we proceeded to build our classifiers. As explained in the Methods section, we compared various classification models, in particular an elastic net with different mixing parameter value (from now called "enet_x" with x being the mixing parameter value), a logistic regression with L1 penalization (named from now "l1") and a logistic regression with L2 penalization (named from now "l2") over the different network feature sets (see Table 1).

For all the combinations, we obtained an accuracy better than a random guess, which would be 0.33 for a 3-class classifier. For each network feature, the accuracies are not statistically different using different penalization types. We also tried a non-linear model, specifically a dense neural network with 5 layers with (128, 64, 256,128,3) neurons, but the results did not improve (maximum accuracy = 0.61 ± 0.04). Based on the results presented in Table 1, we concluded that the different penalization values do not drastically influence the classifier cross-validation performances. Nonetheless, for each network feature we used the penalization maximizing its accuracy on Table 1.

For the text-only classifiers, we observed a higher cross-validated accuracy for AlBERTo (0.67 ± 0.04) than for BERT base (0.52 ± 0.04), following domain specific fine-tuning of the language models on the tweets of the first dataset.

We then paired each of the top-performing models with AlBERTo predictions. The predicted probabilities coming from the text and network classifiers have been combined as in the following:

$$prob = w \cdot prob_{BERT} + (1 - w) \cdot prob_{network}$$

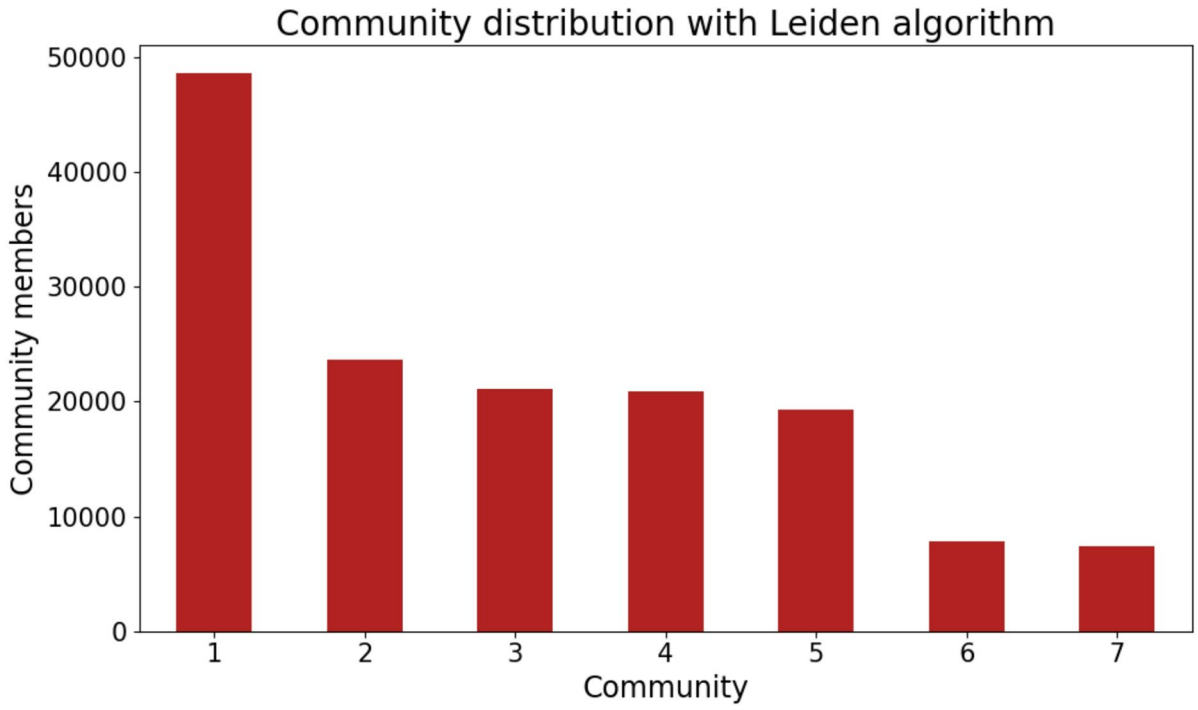

Fig. 1. Community distribution using Leiden algorithm fixing the 90% coverage, The first six communities cover the 90% of the first dataset, while the remaining 10% of the users is placed in the seventh community.

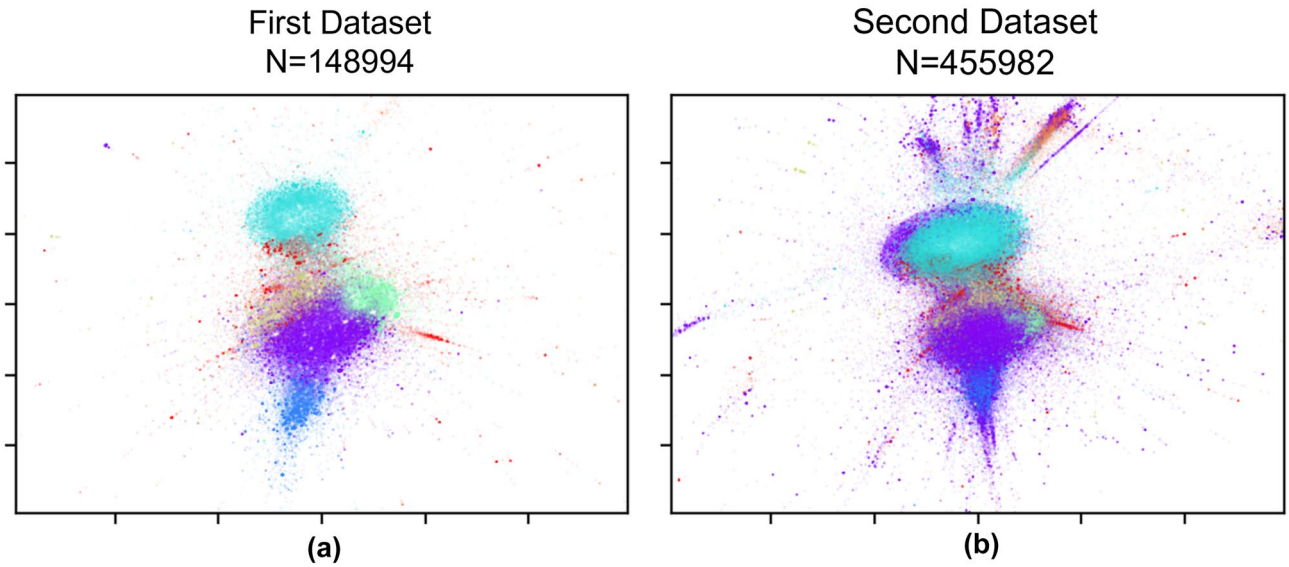

Fig. 2. Network layout using fa2 algorithm. Different colors denote the distinct Leiden communities represented in the visualization. a: first dataset, b: second dataset, N: number of users.

	enet_025	enet_05	enet_075	l1	l2
n2v	**0.631 ± 0.015**	0.630 ± 0.013	0.624 ± 0.011	0.625 ± 0.011	0.629 ± 0.013
fa2	0.62 ± 0.02	0.620 ± 0.019	0.619 ± 0.018	0.619 ± 0.019	**0.62 ± 0.02**
norm_lap	0.613 ± 0.012	0.613 ± 0.012	0.614 ± 0.010	**0.617 ± 0.010**	0.613 ± 0.011
norm_leiden	0.628 ± 0.017	0.628 ± 0.017	**0.630 ± 0.017**	0.629 ± 0.015	0.629 ± 0.017
norm_lab_prop	0.594 ± 0.012	0.596 ± 0.012	0.594 ± 0.013	**0.597 ± 0.014**	0.595 ± 0.014

Table 1. Model cross validation accuracy comparison for each network feature. In bold, the best-performing model per feature. Each row represents a different set of network features: in order node2vec, ForceAtlas 2 coordinates, normalized laplacian eigenvectors and normalized community participation ratios obtained with Leiden algorithm and with label propagation. Each column represents a different classifier: elastic net with different mixing parameters (0.25,0.5,0.75 named enet_025, enet_05 and enet_075), logistic regression with L1 and L2 penalizations.

	n2v + text	fa2 + text	norm_lap + text	norm_leiden + text	norm_lab_prop + text
Model	*enet_025*	*l2*	*l1*	*enet_075*	*l1*
Weights	0,602	0,621	0,607	0,665	0,619
Accuracy	**0.71 ± 0.04**	**0.71 ± 0.05**	0.69 ± 0.04	0.70 ± 0.03	0.70 ± 0.03
f1_score	0.71 ± 0.04	0.69 ± 0.04	0.69 ± 0.04	0.70 ± 0.03	0.70 ± 0.03
Matthews_coefficient	0.57 ± 0.05	0.57 ± 0.056	0.55 ± 0.05	0.55 ± 0.04	0.56 ± 0.04

Table 2. Combination weights for each network feature + text couple. Using AlBERTo's probabilities for the text part and the best model (first row) for the network feature, the accuracies are computed from the cross-validation set. In bold, the best-performing couple text + network feature. We observe very comparable performances and also a similar weight for the text + network combination of classifiers.

with w a weight being fine-tuned with cross-validation over the training set. In Table 2 we show the results from these combination, which were utilized to determine which network feature performed better when associated with the text.

Again, the results demonstrate a similar level of accuracy, with the difference between the best-performing and the worst-performing being approximately 1%, and comparable w weights for the combination of classifiers. Most network features increased their accuracy by about 7–8% when integrated with the text classifier. The network feature characterized by the worst performance when considered alone shows the biggest increase in accuracy (about 12%). Even though fa2 and n2v embeddings performed similarly, we decided to use fa2 + text for the following analysis because it allows better recalculation of embeddings following the addition of new

	n2v	fa2	norm_lap	norm_leiden	norm_lab_prop	text	fa2 + text
Accuracy	0.56 ± 0.03	0.55 ± 0.03	0.42 ± 0.03	0.57 ± 0.03	0.53 ± 0.03	0.52 ± 0.03	**0.62 ± 0.03**
Average_f1_score	0.57 ± 0.03	0.49 ± 0.03	0.34 ± 0.02	0.56 ± 0.03	0.44 ± 0.02	0.52 ± 0.03	**0.61 ± 0.03**
Pro_vax_f1_score	0.60 ± 0.04	0.60 ± 0.04	0.46 ± 0.05	0.57 ± 0.04	0.61 ± 0.04	0.45 ± 0.05	**0.65 ± 0.04**
Neutral_vax_f1_score	0.41 ± 0.04	0.15 ± 0.04	0.03 ± 0.03	0.38 ± 0.05	0.00 ± 0.00	**0.56 ± 0.04**	0.42 ± 0.05
Anti_vax_f1_score	0.69 ± 0.04	0.73 ± 0.04	0.54 ± 0.04	0.72 ± 0.04	0.71 ± 0.04	0.55 ± 0.04	**0.74 ± 0.04**
Matthews_coefficient	0.35 ± 0.04	0.38 ± 0.04	0.18 ± 0.04	0.38 ± 0.04	0.38 ± 0.03	0.28 ± 0.05	**0.47 ± 0.04**

Table 3. Full metrics results on the first test set, each feature's results are computed using the best model for the specific feature. In bold, the best feature per metric. "average_f1_score" is the unweighted average of the 3 per-class f1-scores. "matthews_coefficient" is the Matthew's correlation coefficient for classification.

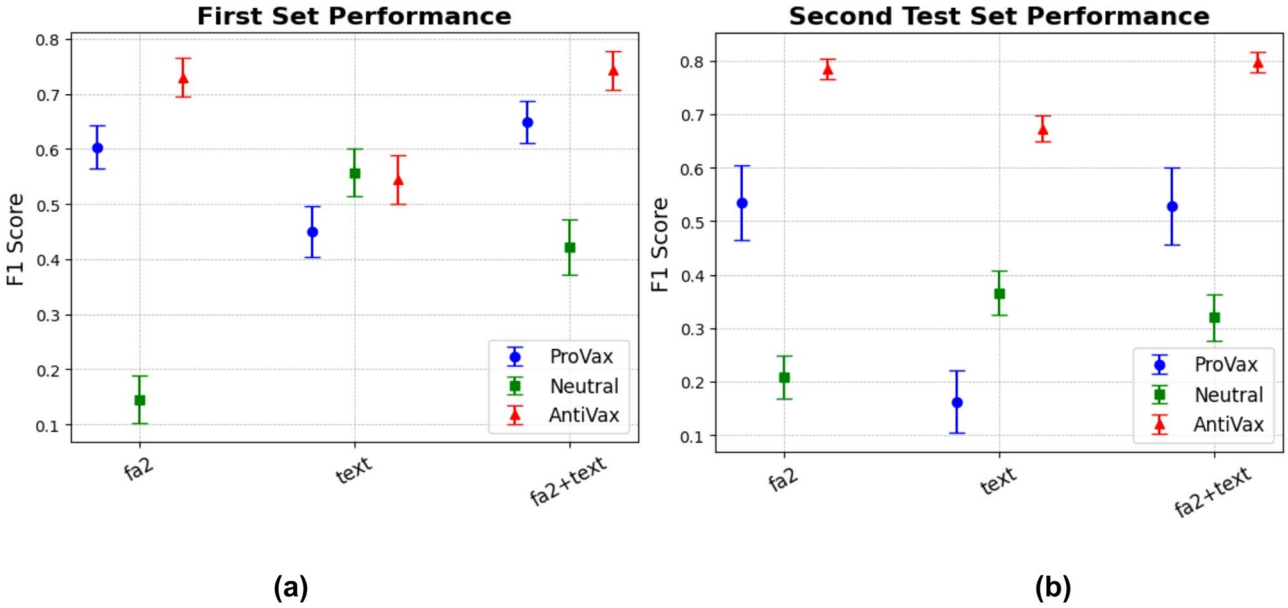

Fig. 3. F1 scores per class in (**a**) first test set and (**b**) in the second dataset, for the integrated classifier (fa2 + text), and classifiers based on fa2 and text separately.

nodes or links. In fact, fa2 allows to use the positions obtained on the first set network as starting points for calculation in the second test network for nodes present in both sets. On the contrary, node2vec needs to be reinitialized from scratch, with no memory of the first test set, with the risk of affecting the classification parameters previously estimated.

For each combination, a slightly higher weight was assigned to the probabilities provided by the text classifier. This choice is probably due to the fact that the text classifier has more balanced probabilities across the three classes. In contrast, the network classifier tends to produce more extreme probability values. For this reason, a balanced combination of the two classifiers requires boosting the probabilities calculated by the text classifier, which otherwise would not affect the final result.

We then evaluated the best model for each feature on the first and second test sets, comparing the various metrics to gain a more comprehensive understanding of the model behavior. In Table 3, we show the performance of each model on the first test set.

The results of the first test set, which is composed of tweets extracted in the same time period as the training set, show a significant performance improvement when combining textual and network information, as opposed to using a single-feature classification approach. Furthermore, we acknowledge that a major challenge in classifying network features arises from the positioning of neutral tweets. This is largely due to the relative ease of network features in distinguishing classes inside highly polarized systems, such as pro-vaccine and anti-vaccine users. As shown in the confusion matrices of Supplementary Figure S2, when network features are included, the classifiers have less chance of misclassifying AntiVax for ProVax and vice versa. The performance of the single network feature is in fact notably superior to that of text classification when identifying ProVax or AntiVax tweets (see Fig. 3; Table 3).

The text classifier demonstrates greater consistency in classifying all three categories (see also Supplementary Figure S2). This leads to poorer outcomes concerning the network classifier when dealing with pro- and anti-vaccine users while showing significantly better performance in classifying neutral users. When both features are

	n2v	fa2	norm_lap	norm_leiden	norm_lab_prop	text	fa2 + text
Accuracy	0.24 ± 0.02	0.64 ± 0.02	0.71 ± 0.02	0.69 ± 0.02	**0.74 ± 0.02**	0.54 ± 0.03	0.66 ± 0.02
Average_f1_score	0.132 ± 0.010	0.51 ± 0.03	0.277 ± 0.005	0.52 ± 0.03	0.46 ± 0.02	0.40 ± 0.03	**0.55 ± 0.03**
Pro_vax_f1_score	0.00 ± 0.00	**0.54 ± 0.07**	0.00 ± 0.00	0.50 ± 0.07	0.51 ± 0.07	0.16 ± 0.06	0.53 ± 0.07
Neutral_vax_f1_score	**0.39 ± 0.03**	0.21 ± 0.04	0.00 ± 0.00	0.23 ± 0.05	0.00 ± 0.00	0.37 ± 0.04	0.32 ± 0.04
Anti_vax_f1_score	0.012 ± 0.009	0.785 ± 0.019	0.832 ± 0.015	0.83 ± 0.02	**0.87 ± 0.01**	0.67 ± 0.02	0.80 ± 0.02
Matthews_coefficient	0.02 ± 0.03	0.23 ± 0.04	0.00 ± 0.00	0.29 ± 0.04	**0.35 ± 0.04**	0.15 ± 0.04	0.30 ± 0.04

Table 4. Full metrics results on the second test set, each feature's results are computed using the best model for the specific feature. In bold, the best feature per metric. "average_f1_score" is the unweighted average of the 3 per-class f1-scores. "matthews_coefficient" is Matthew's correlation coefficient for classification.

combined, the overall effect is averaged, resulting in a still inferior performance on neutral elements compared to the other two categories.

However, there is an improvement of approximately 30% for the F1 score of the worst performing (Neutral) class when comparing classification using fa2 + text versus the results obtained using only fa2.

We then evaluated our model on the second test set (see Table 4), collected and annotated more than a year later. This comparison aims to assess the model's robustness to temporal drift and to determine whether modifications to the network structure and textual approach significantly impact the model's ability to adapt.

On this second dataset, the accuracy of the integrated classifier is comparable to the one computed on the first test set, while we observed a small decrease in average F1-score and Matthew's coefficient. The high accuracy is partially motivated by the imbalance of class labels in the second dataset, which features an increased prevalence of the best-performing class (i.e. AntiVax). Considering the performance of the network features, which we already observed to perform better on polarized classes, we observed a higher accuracy than on the initial dataset (excluding n2v). N2v worse performances are likely explained with the mis-alignment of first and second set embeddings, which is caused by the absence of memory in node2vec computation, as we commented also before. Among the network features, fa2 and norm_leiden have an average F1 score comparable to those in the first test set, and perform relatively better than the other features. The text-only classifier experienced a decline in performance over Neutral and ProVax tweets, as also reflected in the average F1-score, while the F1-score on AntiVax tweets improved, possibly implying that AntiVax users kept their arguments more stable over time, while ProVax and Neutral users likely changed the content of their tweets. As shown in the confusion matrices in Supplementary Figure S3, ProVax and AntiVax tweets are almost fully recognized by network-based classifiers.

Discussion

In this study, we explored the interplay between network features and textual data for the automated classification of user stance on vaccines in online social networks. We collected a Twitter dataset comprising almost 20 million vaccine-related tweets, and we compared the tweet-classification performance of classifiers using text and network features together or separately. We considered all tweets mentioning the "vaccine" topic without any specific preselection, thus, with a more general applicability than other approaches considering only tweets selected through filtering or tagging.

Text-only and network-only classifiers had similar accuracy but with different characteristics, and the integrated classifier over-performed them across all the performance metrics. On the one side, the text classifier helps in recognizing neutral tweets, while network features helped to better classify tweets belonging to the two extreme classes despite being computed at user-level: this suggests that to detect extremal vaccine stances it could be more effective to see *where* a user is placed within the social network rather than *what* they actually write. This result not only confirmed the well-known association between user opinions and social network structure[14,22,33,34], but also showed how to use this knowledge to improve automated stance classification on polarizing topics. From a communication strategy perspective, this result could be particularly valuable, as undecided individuals are often the most responsive to tailored messaging. Targeted interventions, especially when informed by behavioral and psychographic data, have been shown to increase engagement and persuasion effectiveness[42,43].

On the other side, network-based classification remained stable on the second test set, while the text-only classifier suffered a significant drop in performance. This suggests that tweet content and language could drift faster than the topology of the network, possibly needing continuous re-training of the language models as previously observed in similar contexts[44].

In detail, when looking at differences in accuracy for the different classes, on the second dataset the text classifier failed to identify pro-vaccine users, likely due to changes in the content of tweets from users in favor of vaccination at the beginning of the vaccination campaign. By manually inspecting ProVax tweets, we noticed that the users in the first dataset were mostly communicating their intention to vaccinate, and this did not happen anymore in the second period.Moreover, the prevalence of the different classes changed in the second dataset, with ProVax users seeming less prone to participate in the online debate. This imbalance shift toward AntiVax tweets ican be attributed to a waning interest in the topic after the most critical phase of the vaccination process was completed. The users who remained most vocal on the subject were the integralist AntiVax proponents,

which account for the observed disparity. It is, in fact, well-known that vocal minorities are more comfortable expressing unpopular views[45], and this has already been observed in the context of vaccine hesitancy[46].

Conclusion

In this study, we addressed the issue of short text messages classification in social media with respect to their vaccine stance. Our results showed that text-classifiers help to better recognize the intermediate class, which likely differs by tone and content, while network features have the power to maintain a better performance over time, possibly because the social network structure is more stable. Network based features have also proven to be particularly helpful to classify the most extreme classes toward vaccine stance, showing the potential of this approach in the context of highly polarizing topics.

Data availability

Anonymized tweets (no text), network features and manual annotations are deposited in Zenodo: [https://doi.org/10.5281/zenodo.15462619]. The code is hosted on Github at https://github.com/FraDurazzi/TwitterVaccine/tree/tight_deadlines.

Received: 27 May 2025; Accepted: 4 November 2025
Published online: 08 December 2025

References

1. Kaur, S. P. & Gupta, V. COVID-19 vaccine: A comprehensive status report. *Virus Res.* **288**, 198114. https://doi.org/10.1016/j.virusres.2020.198114 (2020).
2. Schultz, É. Ward. Science under Covid-19's magnifying glass: lessons from the first months of the chloroquine debate in the French press. *J. Sociol.* **58** (1), 76–94. https://doi.org/10.1177/1440783321999453 (2022).
3. Garrett, L. COVID-19: the medium is the message. *Lancet* **395** (10228), 942–943. https://doi.org/10.1016/S0140-6736(20)30600-0 (2020).
4. Lorini, C. et al. Measuring health literacy in italy: a validation study of the HLS-EU-Q16 and of the HLS-EU-Q6 in Italian language, conducted in Florence and its surroundings. *Ann. Ist Super Sanita.* **55**, 10–18. https://doi.org/10.4415/ANN_19_01_04 (2019).
5. Kashte, S., Gulbake, A., El-Amin, S. F. & Gupta, I. I. I. A. COVID-19 vaccines: rapid development, implications, challenges and future prospects. *Hum. Cell.* **34** (3), 711–733. https://doi.org/10.1007/s13577-021-00512-4 (2021).
6. Sallam, M. COVID-19 vaccine hesitancy worldwide: A concise systematic review of vaccine acceptance rates. *Vaccines* **9** (2), 160. https://doi.org/10.3390/vaccines9020160 (2021).
7. Browne, M., Thomson, P., Rockloff, M. J. & Pennycook, G. Going against the herd: psychological and cultural factors underlying the 'Vaccination confidence gap'. *PLoS ONE.* **10** (9), e0132562. https://doi.org/10.1371/journal.pone.0132562 (2015).
8. «Behavioural and social drivers of vaccination: tools and practical guidance for achieving high uptake». World Health Organization. Consultato: 3 maggio 2025. [Online]. Disponibile su: https://www.who.int/publications/i/item/9789240049680.
9. Wilson, S. L. & Wiysonge, C. Social media and vaccine hesitancy. *BMJ Glob Health.* **5, 10**, https://doi.org/10.1136/bmjgh-2020-004206 (2020).
10. Puri, N., Coomes, E. A., Haghbayan, H. & Gunaratne, K. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. *Hum. Vaccines Immunother.* **16** (11), 2586–2593. https://doi.org/10.1080/21645515.2020.1780846 (2020).
11. Bonnevie, E., Gallegos-Jeffrey, A., Goldbarg, J., Byrd, B. & Smyser, J. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. *J. Commun. Healthc.* **14** (1), 12–19. https://doi.org/10.1080/17538068.2020.1858222 (2021).
12. Becker, B. F. H. et al. Evaluation of a multinational, multilingual vaccine debate on Twitter. *Vaccine* **34** (50), 6166–6171. https://doi.org/10.1016/j.vaccine.2016.11.007 (2016).
13. Weaver, J. More people search for health online. *NBC News*, 16 luglio 2003. Consultato: 7 marzo 2025. [Online]. Disponibile su: https://www.nbcnews.com/id/wbna3077086
14. Himelboim, I., Xiao, X., Lee, D. K. L., Wang, M. Y. & Borah, P. A social networks approach to Understanding vaccine conversations on twitter: network Clusters, Sentiment, and certainty in HPV social networks. *Health Commun.* **35** (5), 607–615. https://doi.org/10.1080/10410236.2019.1573446 (2020).
15. Featherstone, J. D., Barnett, G. A., Ruiz, J. B. & Zhuang, Y. Millam. Exploring childhood anti-vaccine and pro-vaccine communities on twitter – a perspective from influential users. *Online Soc. Netw. Media.* **20** (100105). https://doi.org/10.1016/j.osnem.2020.100105 (2020).
16. European Centre for Disease Prevention and Control., Systematic scoping review on social media monitoring methods and interventions relating to vaccine hesitancy. LU: Publications Office. Consultato: 17 settembre 2025. [Online]. Disponibile su: https://data.europa.eu/doi/ (2019). https://doi.org/10.2900/260624
17. Mønsted, B. & Lehmann, S. Characterizing polarization in online vaccine discourse—A large-scale study. *PLOS One.* **17** (2), e0263746. https://doi.org/10.1371/journal.pone.0263746 (2022).
18. Hayawi, K., Shahriar, S., Serhani, M. A., Taleb, I. & Mathew, S. S. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public. Health.* **203**, 23–30. https://doi.org/10.1016/j.puhe.2021.11.022 (2022).
19. Kummervold, P. E. et al. Categorizing vaccine confidence with a Transformer-Based machine learning model: analysis of nuances of vaccine sentiment in Twitter discourse. *JMIR Med. Inf.* **9** (10), e29584. https://doi.org/10.2196/29584 (2021).
20. Di Giovanni, M., Corti, L., Pavanetto, S., Pierri, F. & Tocchetti, A. M. Brambilla. A Content-based Approach for the Analysis and Classification of Vaccine-related Stances on Twitter: the Italian Scenario. (2021).
21. Cossard, A., Morales, G. D. F., Kalimeri, K., Mejova, Y. & Paolotti, D. M. Starnini. Falling into the Echo Chamber: The Italian Vaccination Debate on Twitter. In *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 14, pp. 130–140, (2020). https://doi.org/10.1609/icwsm.v14i1.7285.
22. Memon, S. A. K. M. Carley. Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset. *CEUR Workshop Proc.*, vol. 2699, ago. Consultato: 5 febbraio 2021. [Online]. Disponibile su: (2020). http://arxiv.org/abs/2008.00791.
23. European Centre for Disease Prevention and Control, European Food Safety Authority., Tools and methods for promoting vaccination acceptance and uptake: a social and behavioural science approach. LU: Publications Office. Consultato: 17 settembre 2025. [Online]. Disponibile su: https://data.europa.eu/doi/. https://doi.org/10.2900/7701140 (2025).
24. Mitra, T. & Counts, S. J. Pennebaker. Understanding Anti-Vaccination Attitudes in Social Media. *Proc. Int. AAAI Conf. Web Soc. Media* **10**(1), 1. https://doi.org/10.1609/icwsm.v10i1.14729 (2016).
25. Alturayeif, N., Luqman, H. & Ahmed, M. A systematic review of machine learning techniques for stance detection and its applications. *Neural Comput. Appl.* **35** (7), 5113–5144. https://doi.org/10.1007/s00521-023-08285-7 (2023).

26. Kunneman, F., Lambooij, M., Wong, A. & van den Bosch, A. Mollema. Monitoring stance towards vaccination in Twitter messages. *BMC Med. Inf. Decis. Mak.* **20** (1), 33. https://doi.org/10.1186/s12911-020-1046-y (2020).
27. Cheatham, S. et al. Understanding the vaccine stance of Italian tweets and addressing Language changes through the COVID-19 pandemic: development and validation of a machine learning model. *Front. Public. Health*. **10** (1). https://doi.org/10.3389/fpubh.2022.948880 (2022).
28. Vaswani, A. et al. Attention Is All You Need.
29. A. I. Kadhim. Survey on supervised machine learning techniques for automatic text classification. *Artif Intell. Rev*, **52**(1), 273–292. https://doi.org/10.1007/s10462-018-09677-1 (2019).
30. Fields, J., Chovanec, K. & Madiraju, P. A survey of text classification with transformers: how wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access.* **12**, 6518–6531. https://doi.org/10.1109/ACCESS.2024.3349952 (2024).
31. Gasparetto, A., Marcuzzo, M., Zangari, A. & Albarelli, A. A Survey on Text Classification Algorithms: From Text to Predictions. *Information* **13**(2). https://doi.org/10.3390/info13020083 (2022).
32. Gori, D. et al. Mis-tweeting communication: a vaccine hesitancy analysis among Twitter users in Italy. *Acta Bio-Medica Atenei Parm*. **92** (fasc. S6). https://doi.org/10.23750/abm.v92iS6.12251 (2021).
33. Quintana, I. O., Reimann, R., Cheong, M., Alfano, M. & Klein, C. Polarization and trust in the evolution of vaccine discourse on Twitter during COVID-19. *PLOS ONE*. **17** (12), e0277292. https://doi.org/10.1371/journal.pone.0277292 (2022).
34. Dunn, A. G., Leask, J., Zhou, X., Mandl, K. D. & Coiera, E. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. *J. Med. Internet Res.* **17** (6), e144. https://doi.org/10.2196/JMIR.4343 (2015).
35. Faccin, M. Measuring dynamical systems on directed hypergraphs. *Phys. Rev. E.* **106** (3), 034306. https://doi.org/10.1103/PhysRevE.106.034306 (2022).
36. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to leiden: guaranteeing well-connected communities. *Sci. Rep.* **9** (1), 5233. https://doi.org/10.1038/s41598-019-41695-z (2019).
37. Raghavan, U. N., Albert, R. & Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*. **76** (3), 036106. https://doi.org/10.1103/PhysRevE.76.036106 (2007).
38. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONE*. **9** (6). https://doi.org/10.1371/journal.pone.0098679 (2014). e98679.
39. Ch, K. Das. The laplacian spectrum of a graph. *Comput. Math. Appl.* **48** (5), 715–724. https://doi.org/10.1016/j.camwa.2004.05.005 (2004).
40. Grover, A. J. Leskovec. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. New York, NY, USA: Association for Computing Machinery, pp. 855–864. (2016). https://doi.org/10.1145/2939672.2939754.
41. Masoudnia, S. & Ebrahimpour, R. Mixture of experts: a literature survey. *Artif. Intell. Rev.* **42** (2), 275–293. https://doi.org/10.1007/s10462-012-9338-y (2014).
42. Matz, S. C., Kosinski, M., Nave, G. & Stillwell, D. J. Psychological targeting as an effective approach to digital mass persuasion. *Proc. Natl. Acad. Sci. U S A*. **114** (48), 12714–12719. https://doi.org/10.1073/pnas.1710966114 (2017).
43. Milkman, K. L. et al. A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proc. Natl. Acad. Sci.* **118** (20), e2101165118. https://doi.org/10.1073/pnas.2101165118 (2021).
44. [2012. 02197] Addressing machine learning concept drift reveals declining vaccine sentiment during the COVID-19 pandemic. Consultato: 11 settembre 2024. [Online]. Disponibile su: https://arxiv.org/abs/2012.02197.
45. Chaudhry, I. & Gruzd, A. Expressing and challenging racist discourse on facebook: how social media weaken the spiral of silence theory. *Policy Internet*. **12** (1), 88–108. https://doi.org/10.1002/poi3.197 (2020).
46. Blane, J. T. & Bellutta, D. Carley. Social-Cyber maneuvers during the COVID-19 vaccine initial rollout: content analysis of tweets. *J. Med. Internet Res.* **24** (3), e34040. https://doi.org/10.2196/34040 (2022).

Acknowledgements

We thank Daniele Pignedoli for his master's degree thesis which paved the way to this study. We acknowledge the kind contribution of several data annotators: Alexandro Martone, Matteo Grotti, Tommaso Rondini, Isacco Faglioni, Alessia Fideli, Gregorio Berselli, Giacomo Rapparini, Tiziano Latino.

Author contributions

F.D. collected the data and coordinated data annotation. N.B. developed the classification framework. M.F. performed data analysis on the network. D.G. and D.R. supervised the analysis and coordinated the study. All authors wrote and reviewed the manuscript.

Funding

This project was funded through the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874735 "Versatile emerging infectious disease observatory - forecasting, nowcasting and tracking in a changing world (VEO)".

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-27487-8.

Correspondence and requests for materials should be addressed to R.D.