Flipping Stance: Social Influence on Bot's and Non Bot's COVID Vaccine Stance

Lynnette Hui Xian Ng*
CASOS, Institute for Software Research,
Carnegie Mellon University
Pittsburgh, PA, United States
huixiann@andrew.cmu.edu

Kathleen M. Carley
CASOS, Institute for Software Research,
Carnegie Mellon University
Pittsburgh, PA, United States
carley@andrew.cmu.edu

ABSTRACT

Social influence characterizes the change of opinions in a complex social environment, incorporating an individual's past stances and the impact of interpersonal influence through the social network influence. In this work, we observe stance changes towards the coronavirus vaccine on Twitter from April 2020 to May 2021, where 1% of the agents exhibit the stance flipping behavior, of which 53.7% are identified bots. We then propose a novel social influence model to characterize the change in stance of agents. This model considers an agent's and his neighbor's past tweets and the overall network structure towards a stance score. In our experiments, the model achieves 86% accuracy. In our analysis, bot agents require lesser social influence to flip stances and a larger proportion of bots flip.

CCS CONCEPTS

• **Human-centered computing** → **Social network analysis**.

KEYWORDS

Social Influence Model, Stance Prediction

ACM Reference Format:
Lynnette Hui Xian Ng and Kathleen M. Carley. 2021. Flipping Stance: Social Influence on Bot's and Non Bot's COVID Vaccine Stance. In *Proceedings of The Second International MIS2 Workshop: Misinformation and Misbehavior Mining on the Web (MIS2 workshop at KDD 2021).* ACM, New York, NY, USA, 9 pages.

1 INTRODUCTION

Social influence characterizes the change of opinions in a complex social environment, incorporating an individual's static conditions (past posts) and the impact of interpersonal influence from his social network [5]. Previous influence studies on social media involved the construction of an influence locality model to predict retweet behavior using attributes like personal attributes, number of followers/followees and reciprocal following relationships [22], and modelling how majority opinions influence an individual's opinion on Twitter through Markov state transitions [14]. In particular, Xia and Liu observed that an individual's conformity to social influence and initial level of susceptibility are crucial to vaccination stance [20].

The 2020 coronavirus pandemic sent the world into a standstill and researchers scrambled to develop a vaccine that would ease the pandemic. Public opinion about vaccination has always taken two main polarizing camps, pro-vaccine and anti-vaccine. These camps are fondly termed "pro-vaxxers" and "anti-vaxxers", characterized by their stance towards vaccination. Prior work in analyzing the polarizing vaccination debate on social media [9, 18] characterized that both groups exhibit different online behavior in terms of the vaccines discussed, reach and network structure [6]. In general, the two camps interact mostly in separate echo-chambers with the same type of content [16]. Other works investigate another aspect of the debate: personas of state-sponsored actors on the polarization [19, 21] and the higher activity of bots in spreading anti-vaccine messages [3].

In combining social influence and stance detection, previous work ranked Twitter users by social influence in the polarizing BREXIT debate [7], yet others explored the changes in neighborhood overlap of Twitter agent stances [11].

Bringing this combination one step further is the observation of stance changes between decided and undecided in a debate setting, where linguistic factors and audience factors are combined to predict whether an undecided audience member would make a stand [13]. Political studies have also examined the "flip-flopping" of stance in US electoral politics, branding the observation as an attribute of their conviction on issues brought up in presidential debates [12] and other situations like gun-control [2].

Contributions. In this work, we aim to bridge the gap between analyzing the polarizing vaccination debate and identifying individuals susceptible to stance changes due to social influence. We study network and linguistic factors that influence a Twitter agent to flip his stance and propose a novel stance flipping prediction model utilizing social influence to predict stance flips. Our model successfully predicts 86% of the agents whose stance flips in the context of COVID vaccinations. We then analyze the response of bots and non-bots to social influence, observing that bots have less conviction and flip even with fewer neighbors with the opposite stance compared to non-bots. This furthers misinformation research in identifying agents that are susceptible to changing their opinions.

*The research for this paper was supported in part by the Knight Foundation and the Office of Naval Research grant N000141812106 and by the center for Informed Democracy and Social-cybersecurity (IDeaS) and the center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research or the US Government.

MIS2 workshop at KDD 2021, Aug 15, 2021, Virtual

2 DATA AND METHODOLOGY

2.1 Data Collection

We collected Twitter data surrounding the COVID pandemic using the Twitter REST API using the hashtag #coronavirus on a daily basis from 1 April 2020 to 10 May 2021. We begun data collection after the Pfizer-Biotech vaccine began development (in March 2020). We filtered the data to tweets that talk about vaccines, keeping the tweets that have the sub-phrase "vaccine" in one of the hashtags. Additionally, as we are specifically looking for agents that flip stances, we disregard agents that only have 1 tweet in the dataset. Finally, we have 679,235 agents and more than 1.3 million tweets.

2.2 Data Annotation

To describe the tweets further, we labelled the tweets in terms of their stance and linguistic cues. We then combined an agent's tweets to label agents in terms of their overall stance, network centrality and mean linguistic cues.

Bot Annotation. We annotated the data by performing bot-probability annotation using the BotHunter algorithm at the 0.70 threshold level [1]. It extracts account-level metadata and classifies agents using a supervised random forest method through a multi-tiered approach, each tier making use of more features. For each user agent, BotHunter provided a probability that the account is inorganic. A probability over 70% indicates the agent is likely to be a bot. We also annotate users that self-identify as bots through having the word "bot" in their username, i.e. "coronaupdatebot".

Stance Labelling. We manually inspected all vaccine-related hashtags in the dataset and classified the hashtags into pro- and anti-vaccine hashtags. We left out generic hashtags like "#vaccine" and "#covidvaccine" as they do not present a stance. The list is in Appendix 7.2. We use a network-based stance propagation algorithm which models a user-hashtag bipartite graph and propagate the stance labels between the two parts, providing a label and a confidence value for all tweets and agents [10]. We further filter the tweets to those with a defined pro-/anti-stance and their authoring agents.

Linguistic Annotation. Language gives us an insight to an agent's thoughts and emotions[15], and we infer these measures by characterizing linguistic cues. To characterize messages of both groups, we use Netmapper[1] software to count the frequency of key lexical categories including abusive absolutist, positive and negative terms. This builds on psycholinguistic theory associating particular words and expressions with behavioural, cognitive and emotional states [17]. We also included the tweet count as an agent's endogenous variables, as an indication of how expressive the agent is.

Network Annotation. We measure the indication of how an agent is influenced by his neighbors by characterizing social network variables. We used ORA network analysis tool to analyze the network interactions and spread between the agents [4]. We calculated an agent's global centrality values with respect to all the other agents in the entire dataset. The centrality values are: number of followers, eigenvector centrality, total degree centrality, betweenness centrality, super friends and super spreaders. These variables provide an indication of how connected and influential the agents are in the network.

Agent Annotation. Finally, we annotate each agent with their corresponding stance, linguistic and network values, as agents are the focus of the model. We labelled each agent's stance as the stance of his final collected tweet. We additionally kept a chronological history of each agent's stance, which we used in identifying agent stance flipping. We take the mean of each agent's tweets linguistic cues as the agent's overall linguistic cues. Agents' network values are annotated using the values generated from the network annotation, which represents each agent's global centrality value.

3 SOCIAL INFLUENCE ON VACCINE STANCE

In this section, we build a social influence model to evaluate whether an agent would flip his stance towards the vaccine.

3.1 The Social Influence Model

In our social influence model, we describe the formation of a stance towards the coronavirus vaccine (pro-vaccine or anti-vaccine) in terms of an agent's static variables and the interpersonal influences from other agents in the network. We describe an agent's stance in terms of variables and the process linking them. Specifically, an agent's stance towards the vaccine is dependent on his previous stances and linguistic cues of his tweets and his neighbor's information.

Agent stance. We define agent stances Y with the following model: $Y_{agent} = XB$, in which Y is an agent stance outcome score, X is an $1 \times k$ matrix of scores on k endogenous and exogenous variables of the agent and B is a $kx1$ vector of coefficients giving the effects of each of the endogenous variables. In our study, we used agents' linguistic cues as endogenous variables and network values as exogenous variables. Since we are probably analyzing only but a subset of the variables that might affect an agent's stance, we partition the X and B matrices. That is, the equation is modified to represent observations and coefficients of a subset of variables as in Equation 1.

$$Y_{agent} = X_* B_* \tag{1}$$

The Base Influence Model. The base influence model estimates the impact of an agent's past tweets and influence from his neighbors on his stance. Neighbors are other agents that has made communication with the agent in focus. The opinions of these neighbors, or "peers" in the social influence model, have a direct effect on an individual's opinions. On Twitter, this means a reply, retweet or mention by either neighbor agent and agent in focus.

Equation 2 represents the base influence of stance upon an agent by his neighbors. I is an agent's influence stance outcome score, comprising of the sum of stances of the neighbors in the agent's network. A 1st degree neighbor is a node that is one hop away from the agent, a 2nd degree neighbor is nodes two hops away from the agent, and so forth. Based on the number of hops away from the agent, the influence of the node's stance on the agent decreases by a scalar multiple such that each neighbor in that hop contributes an equal influence on the agent in focus. This concept is borrowed from the Katz Centrality concept. For the 1st degree neighbor, each neighbor i of the total n 1st-degree neighbours

[1] http://netanomics.com/netmapper/

contributes $\frac{1}{n}$ influence on the agent; this is further reduced by a scalar multiple of $\frac{1}{m}$ for m 2nd degree neighbors and so on.

$$I = \frac{1}{n}\sum_{i=0}^{n} Y_{\text{1st deg neighbors}} + \frac{1}{n}\frac{1}{m}\sum_{i=0}^{n}\sum_{j=0}^{m} Y_{\text{2nd deg neighbors}} + \frac{1}{n}\frac{1}{m}\frac{1}{l}\sum_{i=0}^{n}\sum_{j=0}^{m}\sum_{k=0}^{l} Y_{\text{3rd deg neighbors}} + \dots \tag{2}$$

Figure 1a illustrates how neighbors are weighted based on their distance to the agent in focus for neighbors up to the 2nd degree. The influence weights w of each neighbor is a function of the number of neighbors the nodes has and the distance to the agent.

We calculated the influence weights of each neighbor as the number of hops from the agent increases for 20,000 agents. The results are plotted in 1b, in which by the elbow rule, the optimal number of hops away from an agent node is 2 hops. The influence per neighbor exponentially decays and tends to 0 as the number of hops increases. As such, our stance flipping prediction model considers only the influence of the first and second degree neighbors. The model that considers only the first degree neighbor influence is the Base Model, presented in Equation 3. Equation 4 extends the base model to evaluate the effect of adding the influence of second degree neighbors. We enhance this base model by adding mechanisms: stance strength, connection and reciprocity.

$$I = \alpha\left[\sum_{i=0}^{n} Y_{\text{1st deg neighbors}}\right], \text{where } \alpha = \frac{1}{n} \tag{3}$$

$$I = \alpha\left[\sum_{i=0}^{n} Y_{\text{1st deg neighbors}} + \sum_{i=0}^{n}\sum_{j=0}^{m} \beta Y_{\text{2nd deg neighbors}}\right], \text{where } \alpha = \frac{1}{n}, \beta = \frac{1}{m} \tag{4}$$

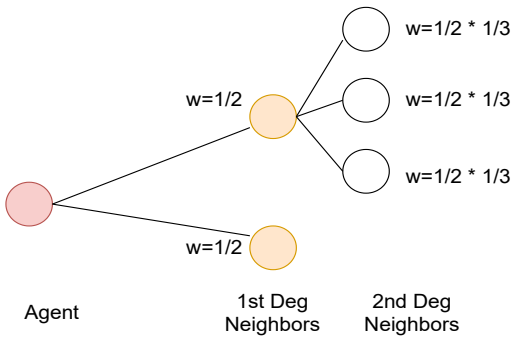
Stance Strength. The first mechanism we add is the effect of stance strength on an agent's outcome score: $Y_{agent} = \gamma X_* B_*$, where γ is a scalar representing the agent's stance strength and its importance. Stance strength alludes to the fact that the more an agent expresses a stance, the stronger he believes in it. It is defined as the proportion the final stance s_{final} is expressed against the number of expressed stances s, multiplied by the a variable importance value w_s. With this mechanism, neighbor's stances Y are calculated similarly.

$$\gamma = \frac{|s_{final}|}{|s|} \times w_s \tag{5}$$

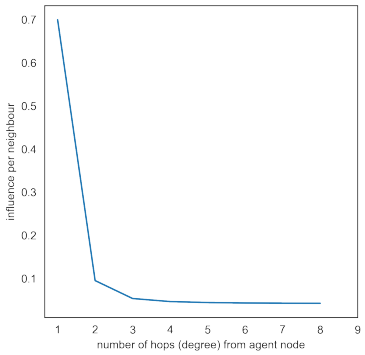
Connection. Connection C is the proportion of neighbors that support an agent's stance:

$$C_{agent} = \frac{\text{\#neighbors with same stance}}{\text{\#neighbors}} \tag{6}$$

Connection represents opinion similarity between the agent and his neighbors, which lends strength to the stance an agent expressed. Collectively, agent (and neighboring agent) stances are enhanced with this mechanism: $Y_{agent} = \gamma C X_* B_*$.



(a) Illustration of the calculation of influence weight w of each neighbor, taking in account their distance from the agent in focus.



(b) Influence weight of each neighbor tends to 0 as the number of hops from the agent increases.

Figure 1: Illustration of influence weights calculation and neighbor plots

Reciprocity. Reciprocity R is the two way interaction between two agents. The higher the reciprocity value, the closer the agents are in friendship, leading to a higher influence on the agent.

$$R = 2 \times \text{\#reciprocal interactions} \tag{7}$$

This mechanism thus modifies the stance score of a neighbor agent: $Y_{neighbor} = \gamma C X_* B_* + R$. The stance score of the agent in focus remains the same.

Deflection Score. We define a deflection score in Equation 8, which characterizes the difference in the score between the agent's stance and the influences from the variables. The agent i will flip his stance if $D_i \geq \epsilon$. For the base model, we set ϵ at 10% of the number of agents. For the models with 2nd degree neighbor information, ϵ is 1% of the total number of agents, reflecting the proportion of the number of agents that flip stances in the overall dataset.

$$D = (I - Y_{agent})^2 \tag{8}$$

3.2 Determining variable importance

We need to determine the coefficients B_* of the variables in the model. As such, we performed a binary classification task with a decision tree model using the Python sklearn library. We run this decision tree across the entire dataset to collectively determine

feature importances. The task uses all the defined linguistic and network variables to classify whether the agent flips or not. We performed a five-fold cross-validation with an 80-20 train-test split. To account for the huge class imbalance, we used the stratified sampling method which makes sure that both the train and test sets have both types of agents. We then extracted the feature importance from the decision tree model, which are used as the variable importance matrix B_*.

3.3 Experimental setup

We apply the social influence model on our dataset to predict stance flips. We only investigate agents who have more than 1 tweet in order to have changes in vaccine stances. For these agents, we leave out each agent's last stance, and use the collected historical data to predict the final stance. However, in the collected historical data, we do include agents that have only one tweet, as they contribute influence to the agents in focus. We progressively add mechanisms to the base model, studying the effects each variable has on improving the model. We measure the macro-F1 accuracy to factor for the unbalanced dataset. Then, we analyze the social influence and tendency to flip by the two classes of data: bots and non-bots.

4 RESULTS

4.1 Summary of Data

Across the dataset, we collected 679k agents with 1.3 million tweets. The vaccine-related tweets were mostly of the languages English, French and Spanish. 32% of the dataset are classified as bots and 1.6% of the agents self-identify as bots. The proportion of stances for tweets and agents in the dataset are around the same: 90% provaccine and 10% anti-vaccine. In total, only 1% of the agents exhibited the stance flipping behavior. Most agents flipped from provaccine to anti-vaccine. Table 1 shows two examples of agents that flipped from pro-vaccine to anti-vaccine stances. These are original messages, i.e. the messages are written by the agents themselves and are not retweets, quoted tweets or replies.

4.2 Social Influence on COVID Vaccine Stance

Based on a five-fold cross validation run on a decision tree, the most important features are: (a) linguistic variables: number of tweets, average word and sentence length, reading difficulty; (b) network variables: number of followers, eigenvector centrality, super spreaders and betweenness centrality. These importance values are used as the coefficients in B_* in the social influence model. The importance scores of each endogenous and exogenous variables of the agents are reflected in the Appendix at Table 4. The importance values are used as the coefficients in the social influence model.

We perform incremental experimental runs on our dataset, each run adding a mechanism to the model. The results are presented in Table 2. Our final stance flipping model outperforms all the other models with an accuracy score of 86%, showing that a combination of all the identified factors are important to the influence of the agent stances. Ablation analysis where we removed either network or linguistic variables in the base model shows a low prediction score of around 0.17%, indicating that both linguistic and network variables contribute to the success of the model prediction.

Our baseline decision tree model performs at 37% accuracy. Our base prediction model that takes in only the first degree neighbor information performs similar to the decision tree model. The accuracy increases 11% with the addition of information from second degree neighbors, indicating the importance of indirect influence on an agent stance. While there is a slight 5% increase in accuracy with the addition of connection, the accuracy increases drastically at 11% with addition of reciprocal ties, showing that the stronger the tie between two agents, the stronger the influence mechanism.

4.3 Bot and Non-bot agents

Out of the agents that flip their stances, 53.7% are identified as bots by the BotHunter algorithm. 6.6% of the overall bot population flip stances while only 2.7% of non-bot agents flip stances. Bots are easier to predict, resulting in a higher accuracy score than non-bots agents. We show an example of an identified bot agent that flip stance in Table 3. This account repeats a message from the anti-vaccine camp several times, before repeating a message from the pro-vaccine camp.

The bot population has a lower deflection score than the non-bot population and the overall population average, which is visualised in Figure 4. The histogram of deflection scores of non-bots are shifted to the right, showing they generally form more interactions with other agents (connections/ reciprocal) and are more convicted on their stance. Figure 2 and 3 show positive results for bot and non-bot agents respectively: agents that are predicted to flip according to the social influence model and their final stance indeed is a flipped stance. In general, we observe that agents that are detected to flip have a very strong network influence of the opposite stance, emphasizing the importance of peer effects, where connected agents have a strong influence over an agent's opinion. Compared to non-bot agents that flip stances (Figure 3, non-bot graphs are typically very sparse and connected to one or two other large clusters. Self-declared bot accounts with the word "bot" in their user names typically have large deflection scores that are in the 95th-percentile zone of the deflection scores dataset, and 5% of these agents flip stances.

Figures 5a and 5b depict the deflection scores against the number of 1st and 2nd degree neighbors that have the opposite stance. Bot agents flip stances at a lower value of opposite stances of neighbors compared to non-bot agents, i.e. the number of 1st and 2nd degree neighbors with stances opposite to the agent's current stance. Non-bot agents are harder to predict as there is no clear distinction between the deflection scores of agents that flip and those that don't.

5 DISCUSSION

In this paper, we constructed a social influence model to predict whether an agent on Twitter will change his stance towards the coronavirus vaccination. The model was incrementally built from the base influence model which estimates the impact of an agent's past tweets and his neighbors on his stances, then additional mechanisms of stance strength, connection and reciprocity were added. We also further investigate whether social influence has differences between bot and non-bot accounts.

Agent 1	Agent 2
If you're vaccine hesitant, just a reminder: COVID-19 is not remotely human hesitant #COVID19 #VaccinesWork	When a business has a 20 times return on investments u push for it the best u can #business #VaccinesWork #covid19
There is no way a vaccine can be dmonstrated to be safe and ready before year's end. I would love to be wrong on this	#CovidVaccine #VaccinePassport #COVIDIOTS #covid19 why risk your precious health on a trial vacc for a disease with over 97% recovery & higher for young people
Industry leaders release an open letter calling on #COVID19 #vaccine makers & govt for honesty [..]	Rachel's family have reported her death as a 'yellow card' as she developed her symptoms after receiving a coronavirus vaccine [...]

Table 1: Examples of original tweets showing agent stance flips. Black and red text are pro and anti-vaccine texts respectively.

Model #	Model	Accuracy
Baseline	Decision Tree	0.38
Model 1	Base social influence model	0.37
Model 2	Model 1 + 2nd deg neighbor information	0.48
Model 3	Model 2 + stance strength	0.70
Model 4	Model 3 + connection	0.75
Model 5	Model 4 + reciprocity	0.86
Ablations		
Model 1 - network	Base social influence model without network variables	0.17
Model 1 - linguistic	Base social influence model without linguistic variables	0.19
Bots only	Model 5 with only bot agents	0.73
Non-Bots only	Model 5 with only non-bot agents	0.67

Table 2: Results of Social Influence Models. The base social influence model is defined in Equation 3.

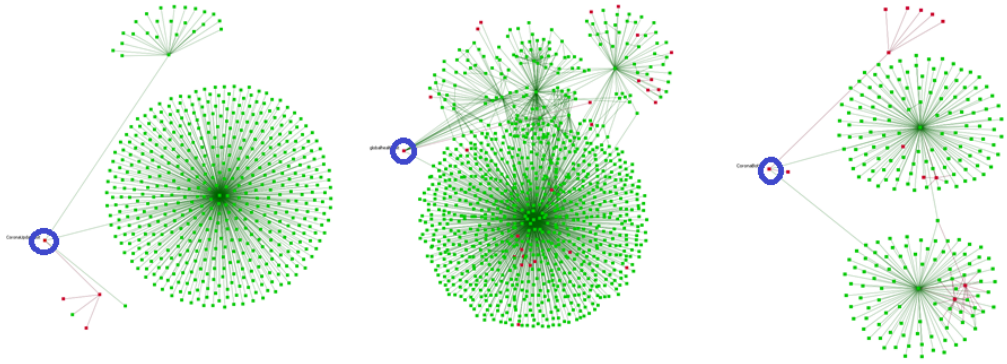


Figure 2: Network graphs of bot agents that our model correctly predicts to flip. Green and red nodes are pro- and anti-vaccine agents respectively. The color of the agent stance is the stance before the flip.

In our estimate of linguistic variable importance, the variables word length, sentence length and Flesch–Kincaid reading difficulty score relates to readability of the tweets. Tweets that are easier to read catches other agents attention better. We observe that more importance is placed on 2nd and 3rd person pronouns compared to 1st person pronouns. Pronouns highlight the attention of the author [8]– 2nd person pronouns like "you" directly addresses the reader and pulls him closer to the author; 1st person pronouns like "we" signifies the authors as embedded within a social relationship, making for a more inclusive conversation; 3rd person pronouns like "she/he" expresses opinion of others as a distinct identity from the author. In our estimate the network variable importance, we identify that eigenvector and betweenness centrality has a high variable importance. These two measures signify the influence an agent has in a network, based on the concept of connections to influential agents and information flow respectively. An agent's position in the social network is one of the key factors in influencing others.

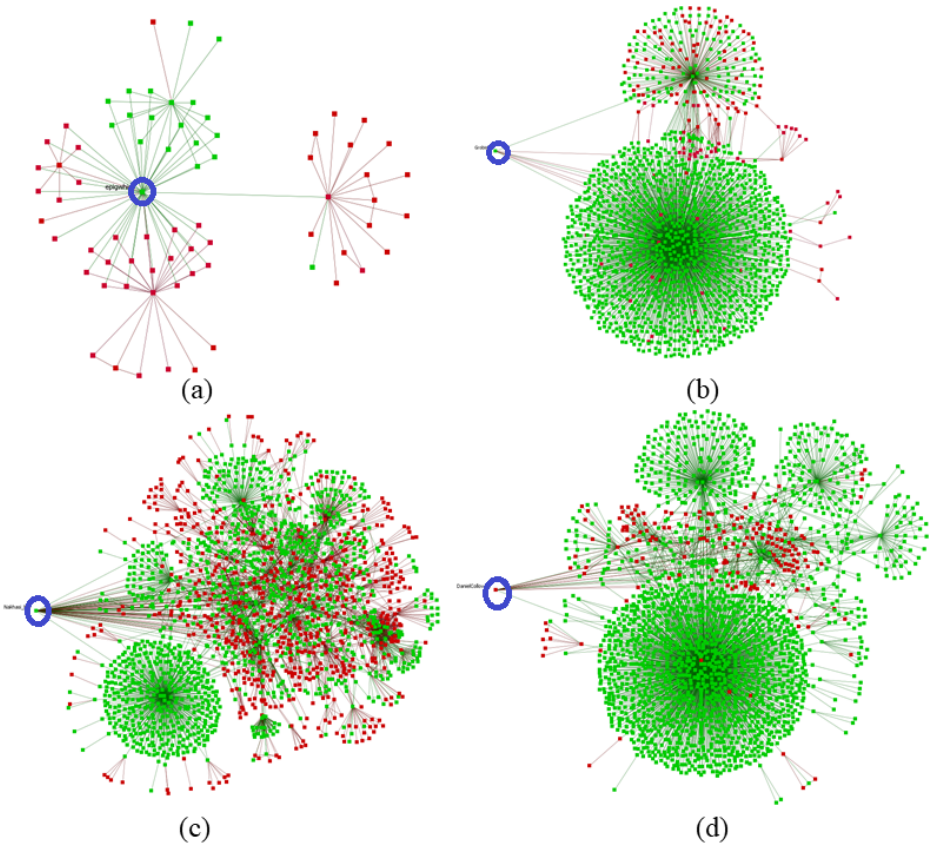


Figure 3: Network graphs of non-bot agents that our model correctly predicts to flip. Green and red nodes are pro- and anti-vaccine agents respectively. The color of the agent stance is the stance before the flip.

Stance	Tweet
antivax	DO NOT TAKE THE #COVID VACCINE #VaccineInjury #VaccineDamage #covidHOAX *message repeated several times*
antivax	Mum passed away after taking experimental vaccine [...] *message repeated several times*
provax	"Safe and effective" #coronavirus #Covid_19 #CovidVaccine #VaccinesWork #vaccinated *message repeated several times*
provax	I was proud to get the COVID-19 vaccine earlier today at Morehouse School of Medicine. I hope you do the same!

Table 3: Sample messages of a bot agent that flip stances

Our results contribute to the reflection of factors that influence an agent's stance in online social media: a combination of network and linguistic variables is crucial in predicting an agent's future stance. An agent is deeply influenced by the opinions of the network of neighbors around him, as observed from the increased in accuracy after the addition of second degree neighbor information and reciprocal ties. In addition, one's conviction towards a stance plays an important role in the agent flipping behavior. In our model,

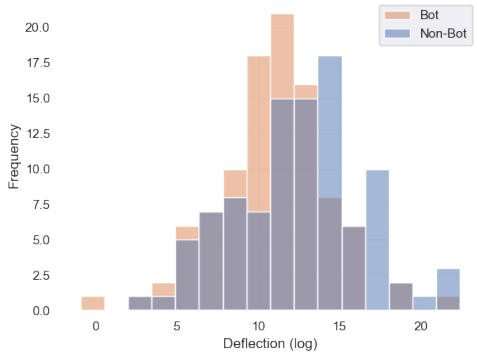
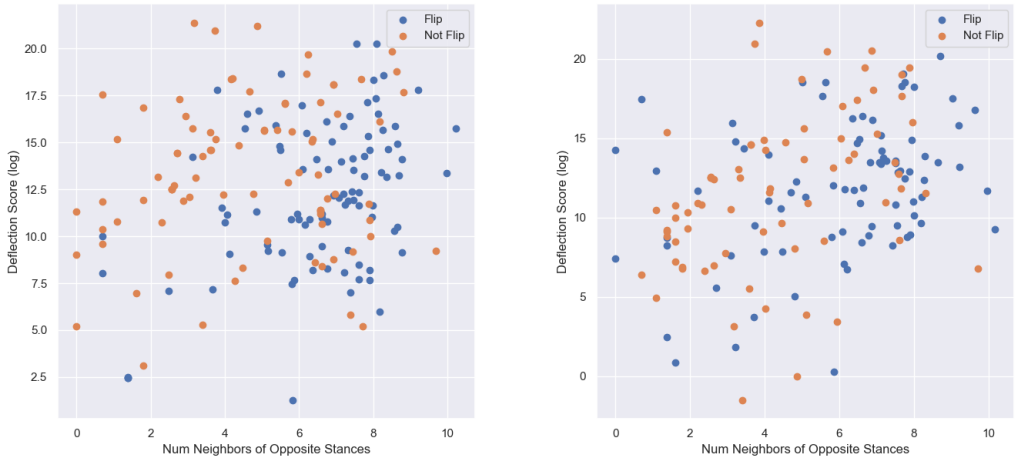


Figure 4: Histogram of deflection scores (log scale) for bot and non-bot populations.

this is represented by stance strength, which is an indication on how easily an agent can be influenced.

In our contrast analysis of bots and non-bot agents, we observe that bot agents have lesser conviction and a larger proportion flip stances (6.6%). Bot accounts flip even with fewer neighbors in the opposite direction of stances. In contrast, non-bot accounts have more conviction and a smaller proportion flip (2.7%) flip stances and require more neighbors of the opposite stance to flip. For accounts



(a) Bots stances **(b) Non-Bots stances**

Figure 5: Deflection plots of Bots and Non-Bot agents

that declare themselves as bot accounts by the use of the word "bot" in their account name, the proportion of these agents flipping stances is five times higher than the population proportion. Agents in this group that flip stances tend to have large deflection scores, signifying that they mix with communities that are predominantly different from their original stance.

Bot accounts typically repeat the same message from one stance several times before switching stance and repeating another message. While intuitively we expect mis/disinformation bots to hold firm to their stance and not be impacted by influence from neighbors, we postulate that bot agents easily flip their stances to match their neighbors' stances, possibly to fit into their surrounding network, so their future tweets have a higher chance of getting viewed by the network. This involves further investigation from a longitudinal perspective.

Bot agents are also easier to predict stance flips, as observed by the higher accuracy score compared to the non-bot group. This, together with bots requiring a lower number of neighbors of the opposite stance for a flipped stance, suggests that bots are more prone to flipping their stances.

Several limitations nuance our conclusions from this work. Users with extreme opinions are typically more vocal on social media, suggesting caution in extrapolating findings. The list of pro-vaccine and anti-vaccine hashtags needs to be continually updated as new hashtags emerge in social media lingo. Nonetheless, we hope that our work provides an understanding into characterizing agents who flip their stances on Twitter, and provides an insight into the difference difference in influence bot and non-bot agents require to change its stance. In future work, we hope to incorporate apriori assumptions about content like an agent's personal values in our stance flipping model and experimenting with a graph neural network model.

6 CONCLUSION

In this study, we observe stance changes towards the COVID vaccine on Twitter from April 2020 to May 2021, where 1% of the agents exhibit the stance flipping behavior. To predict stance changes, we propose a novel model of stance dynamics in the Twitter social network which integrates linguistic information from an agent's past tweets and interpersonal influence from an agent's network connection. The model predicts whether an agent will flip his stance with 86% accuracy. In a contrast analysis between bots and non-bot agents, we identify that a larger proportion of bot agents flip and they flip even with fewer neighbors in the opposite direction of stances, signifying the social influence on these agents can be lesser compared to non-bot agents for them to change their stance.

REFERENCES

[1] David M Beskow and Kathleen M Carley. 2018. Bot-hunter: a tiered approach to detecting & characterizing automated activity on twitter. In *Conference paper. SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, Vol. 3. 3.

[2] Laurent Bouton, Paola Conconi, Francisco Pino, and Maurizio Zanardi. 2014. *Guns and votes.* Technical Report. National Bureau of Economic Research.

[3] David A. Broniatowski, Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. 2018. Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *American Journal of Public Health* 108, 10 (2018), 1378–1384. https://doi.org/10.2105/AJPH.2018.304567 arXiv:https://doi.org/10.2105/AJPH.2018.304567 PMID: 30138075.

[4] L Richard Carley, Jeff Reminga, and Kathleen M Carley. 2018. ORA & NetMapper. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer.

[5] Noah E. Friedkin and Eugene C. Johnsen. 1990. Social influence and opinions. *The Journal of Mathematical Sociology* 15, 3-4 (1990), 193–206. https://doi.org/10.1080/0022250X.1990.9990069 arXiv:https://doi.org/10.1080/0022250X.1990.9990069

[6] Floriana Gargiulo, Florian Cafiero, Paul Guille-Escuret, Valérie Seror, and Jeremy K. Ward. 2020. Asymmetric participation of defenders and critics of vaccines to debates on French-speaking Twitter. *Scientific Reports* 10, 1 (20 Apr 2020), 6599. https://doi.org/10.1038/s41598-020-62880-5

[7] Miha Grčar, Darko Cherepnalkoski, Igor Mozetič, and Petra Kralj Novak. 2017. Stance and influence of Twitter users regarding the Brexit referendum. *Computational Social Networks* 4, 1 (24 Jul 2017), 6. https://doi.org/10.1186/s40649-017-0042-6

[8] Ewa Kacewicz, James W. Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C. Graesser. 2014. Pronoun Use Reflects Standings in Social Hierarchies. *Journal of Language and Social Psychology* 33, 2 (2014), 125–143. https://doi.org/10.1177/0261927X13502654 arXiv:https://doi.org/10.1177/0261927X13502654

[9] Gloria J Kang, Sinclair R Ewing-Nelson, Lauren Mackey, James T Schlitt, Achla Marathe, Kaja M Abbas, and Samarth Swarup. 2017. Semantic network analysis of vaccine sentiment in online social media. *Vaccine* 35, 29 (2017), 3621–3638.

[10] Sumeet Kumar. 2020. *Social media analytics for stance mining a multi-modal approach with weak supervision.* Ph.D. Dissertation. Carnegie Mellon University.

[11] Mirko Lai, Marcella Tambuscio, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2017. Extracting Graph Topological Information and Users' Opinion. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Gareth J.F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro (Eds.). Springer International Publishing, Cham, 112–118.

[12] Michael Lempert. 2009. On 'flip-flopping': Branded stance-taking in US electoral politics 1. *Journal of Sociolinguistics* 13, 2 (2009), 223–248.

[13] Liane Longpre, Esin Durmus, and Claire Cardie. 2019. Persuasion of the Undecided: Language vs. the Listener. In *Proceedings of the 6th Workshop on Argument Mining*. Association for Computational Linguistics, Florence, Italy, 167–176. https://doi.org/10.18653/v1/W19-4519

[14] Debashis Naskar, Sanasam Ranbir Singh, Durgesh Kumar, Sukumar Nandi, and Eva Onaindia de la Rivaherrera. 2020. Emotion dynamics of public opinions on twitter. *ACM Transactions on Information Systems (TOIS)* 38, 2 (2020), 1–24.

[15] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.

[16] Ana Lucía Schmidt, Fabiana Zollo, Antonio Scala, Cornelia Betsch, and Walter Quattrociocchi. 2018. Polarization of the vaccination debate on Facebook. *Vaccine* 36, 25 (2018), 3606–3612. https://doi.org/10.1016/j.vaccine.2018.05.040

[17] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54.

[18] Francois van Schalkwyk, Jonathan Dudek, and Rodrigo Costas. 2020. Communities of shared interests and cognitive bridges: the case of the anti-vaccination movement on Twitter. *Scientometrics* 125, 2 (2020), 1499–1516.

[19] Dror Walter, Yotam Ophir, and Kathleen Hall Jamieson. 2020. Russian Twitter Accounts and the Partisan Polarization of Vaccine Discourse, 2015–2017. *American Journal of Public Health* 110, 5 (2020), 718–724. https://doi.org/10.2105/AJPH.2019.305564 arXiv:https://doi.org/10.2105/AJPH.2019.305564 PMID: 32191516.

[20] Shang Xia and Jiming Liu. 2013. A Computational Approach to Characterizing the Impact of Social Influence on Individuals' Vaccination Decision Making. *PLOS ONE* 8, 4 (04 2013), 1–11. https://doi.org/10.1371/journal.pone.0060373

[21] Xiaoyi Yuan, Ross J. Schuchard, and Andrew T. Crooks. 2019. Examining Emergent Communities and Social Bots Within the Polarized Online Vaccination Debate in Twitter. *Social Media + Society* 5, 3 (2019), 2056305119865465. https://doi.org/10.1177/2056305119865465 arXiv:https://doi.org/10.1177/2056305119865465

[22] Jing Zhang, Jie Tang, Juanzi Li, Yang Liu, and Chunxiao Xing. 2015. Who influenced you? predicting retweet via social influence locality. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 9, 3 (2015), 1–26.

7 APPENDIX

7.1 Variable Importance

Table 4 refers to the importance of endogenous and exogenous variables of the agents. The importance values are used as the coefficients in the social influence model.

Variable	Importance
Linguistic Variables	
Tweet count	0.224
Avg word length	0.068
Reading difficulty	0.054
Positive sentiment	0.040
Negative sentiment	0.029
Agent's own stance	0.026
Num identities terms	0.026
Num pronouns used	0.021
2nd person pronouns	0.021
3rd person pronouns	0.018
Num exclamation points	0.015
1st person pronouns	0.015
Num family terms	0.009
Num exclusive terms	0.008
Num abusive terms	0.007
Network Variables	
Num of followers	0.194
Eigenvector centrality	0.087
Super spreaders	0.043
Betweenness centrality	0.033
Super friends	0.022

Table 4: Variable importance from the decision tree model

7.2 Stance-related hashtags

The following are the hashtags used for annotating the pro- and anti- vaccine stances.

Pro-Vaccine hashtags: VaccinesWork, Sharethevaccine, ProtectVaccineProgress, getvaccine, WaitforVaccine, FreeTheVaccine, vaccinesaresafe, vaccineconfidence, igotvaccinated, coronavaccineforall, CoronavirusVaccineAppointment, justtakethefuckingvaccine, VaccinesWithoutBorders, Vaccines4All, vaccinesaves, Vaccine4ALL, BreakthroughVaccine, waitingformyvaccine, GetTheVaccine, GetYourVaccine, takethevaccine, vaccinesafe, Vaccineswillwork, Iwilltakethevaccine, vaccinefreedom, VaccineToAll, SafeAndEffectiveCovid19Vaccine, VACCINESARESAFE, safevaccines, nosleeptilvaccine, NoOnsiteSchoolsUntilVACCINES, WhereIsMyVaccine, vaccineselfie, vaccineready, vaccinesareamazing, vaccinee, VirusToVaccine2020, HoHoHopeVaccineArrivesSoon, Vaccined, VaccinesWork, VaccinesSaveLives, Vaccine4All, effectivevaccine, VaccineForSA, VaccinesSavesLives, SafeVaccines, accesstovaccines, VaccineHope,

VaccinesWithoutBorders, wherearethevaccines, VaccinesforAll, getthevaccine, giveusthevaccine, VaccineWorks, vaccinesavelifes, GetAVaccine, vaccinesafelife, safevaccine, HaveTheVaccine, WhyIGotMyVaccine, CovidVaccineforall, vaccinesavestheworld, ImGettingTheVaccine, FreeVaccines, VaccinesWorkforAll, GiveMeMyVaccineNow, Affordablevaccine4all, getthatvaccine, justgivemethevaccine, SayYestotheVaccine, TakeYourVaccine, provaccine, YesToVaccine, vaccinesave, covid19_vaccine_4all, vaccineissafe, PleaseGetTheVaccine, VaccinesForEveryone, TrustTheVaccine, VaccinesSaveLives, havevaccinewilltravel, VaccinesWork, ArmysForVaccines, VaccinesForAll, VaccineForAll, We4Vaccine, vacciner, TakeTheVaccine, affordablevaccine4all, Votes4Vaccine, stopvaccinepolitics, SafeVaccines, NeedVaccine, GetTheVaccine, TrustYourVaccine, YesToCovid19Vaccine, WinWithVaccine, VaccineDay, VaccineSelfie, WheresTheVaccine, HaveTheCovidVaccine, coronavaccineforall, CovidVaccineToday, SecondVaccine, VaccineisSafe, vaccinesafetyadvocate, VACCINEROUNDTHECLOCKNOW, igotmycovid19vaccine, VaccineWorks, vaccineacceptance, VaccinesAreSafe, getmorevaccine, weneedthevaccine, TeamVaccines, TeamVaccine, GoForVaccine, getVaccine, provaccine, India4Vaccine, vaccinesmatter, VaccineNow, vaccinesavelivess, VaccineFTW, HaveTheVaccine, GetYourVaccine, vaccinesaveslives, VaccineWork, IWantMyVaccine, ProVaccines, vaccinesareamazing, makevaccinesfree, vaccine4all, rolloutthevaccine, WhereIsMyVaccine, VaccineSavesLives, TrustTheVaccine, covid19vaccineyesplease, getthecovidvaccine, covid19vaccine4all, SayYesToVaccine, ineedthevaccine, GetUsVaccines

Anti-Vaccine hashtags: NoVaccines, NoVaccinesForMe, VaccineYourAss, NoToCoronavirusVaccines, SayNoToVaccine, NoMandatoryVaccine, VaccinesKill, VaccinesKills, stopvaccine, fkyourvaccines, antivaccine, AntiVaccine, NoVaccine, StopCovidVaccine, WeDoNotConsentCVVaccine, VaccinesKill, VaccinesHarm, SayNoToVaccines, ForcedVaccines, VaccineIsPoison, ResistVaccines, Noneedvaccines, FakeCoronaVaccine, VaccinesAreBioweapons, Iwontgetthevaccine, VaccineFromHell, vaccinepoison, StopAllVaccines, Vaccinetakedown, vaccinesDAMANGEimmunity, vaccinesRnotNATURAL, RejectTheVaccines, BewareVaccines, FuckVaccines, HellNoVaccine, NoVaccinesEver, WhoNeedsVaccine, justsaynotoforcedvaccines, StickTheVaccineUpYourArse, nottakingavaccine, vaccinebad, WeaponizedDeadlyVaccines, JustSayNOToTheVaccines, DoNotTakeTheVaccine, FuckVaccinePoison, NOVaccine4Me, VaccinesNotSafe, VaccineBioWeapon, wedontwantvaccine, vaccinenotforme, DontTakeCovidVaccine, NotTakingCovidVaccine, DangerousVaccine, vaccinedoesnotwork, VaccineIsntSafe, No2Vaccine, OpposeTheVaccine, YouCanHaveMyVaccine, ShoveThatVaccineUpYourAsshole, MoThankYouCovidVaccine, ToHellWithCovidVaccine, CovidVaccinePoison, SCREWTHEVACCINE, murdervaccine, VaccineIsUseless, KilloronaVaccine, DodgyVaccine, DoNotTakeAnyVaccine, JeNeMeVaccineraiPas, NoVaccineForMe, stopthevaccine, novaccine, BoycottIndianVaccine, vaccinehesitancy, To_Vaccine_Is_My_Choice, vaccinefailure, donttakethevaccine, FakeVaccine, StolenVaccines, NoVaccine4Me, TrudeauVaccineContractsLie, vaccinesKill, VaccinesArePoison, VaccinesAreNotCures, NotAVaccine, jesusisvaccine, HALTtheVaccines, NoVaccines, dontgetthevaccine, VaccinesHarm, TeamNoVaccine, TheVaccineIsTheVirus, fakeVaccines, killervaccine, lethalvaccine, CoronaVaccineFail, vaccineBioweapon, saynotovaccines, Fuckvaccines, fuckyourvaccines, vaccinedeath, JustSayMoToVaccines, NoCOVIDVAccineMandate, NoMandatoryVaccines, GoHomeLeaveOurVaccinesAlone, antivaccine, anti_vaccine, stopcovidvaccine, CovidVaccineHesitancy, VaccinesCanKill, Antivaccines, notovaccine, DontGetVaccine, VaccineDontWork, VaccineKills, StopVaccine, ForcedVaccine, FakeCovidVaccine, VaccinesAreBad, Falsevaccine, PfizerVaccineKillingPeople, NoCovidVaccineForMe, VaccineShaming, TakeTheVaccineAndShoveItUpYourAss, FakeVaccinesWillNotSaveYou, WorldSaysNoVaccine, NOCovidVaccine, NOCovid19Vaccine, Jesusovervaccines, IAmNOTVAccineBait, NotAVaccineAMedicalExperiment, VaccineMortality, NoVaccine, NoVaccineForMe, NOVACCINE4ME, VaccineDeaths, VaccineDeath, VaccineInjury, AntiVaccine, vaccinesideeffect, NoToCoronaVirusVaccines, AvoidCovidVaccine, NoVaccines, coronavirusvaccinescam, ForcedVaccines, destroyvaccines, VaccineHesitancy, noneed4vaccine, notocovidvaccine, vaccinesharm, vaccinedamage, vaccinesareevil, SayNoToVaccines, killervaccine, no2vaccine, vaccinekills, vaccineskill, VaccinesKillingpeople, JustSayNotoVaccines, vaccinedanger, donttrustthecovid19vaccine, RejectWeaponizedVaccines, StopVaccine, FakeVaccine, vaccinechemicalweapon, DeathToVaccines, VaccineScam, NoVaccinesNeeded, NoVaccinesForMe, vaccinehesitant, PoisonVaccine, DeathVaccine, Cancervaccine, VaccineFail, vaccineRESISTANT, VaccineNonsense, UnsafeVaccines, NoVaccinesNecessary, AbleanableVACCINE, Vaccinefuckup, novaccinerequired, vaccinedrivemutations, VaccineFromHell, GodIsMyVaccine, AntiCovidVaccine, NeitherDoTheseCOVIDVaccines, notovaccine, Notmyvaccine, CovidVaccineIsPoison, VaccineDisaster, NovavaxVaccine, TheCovidVaccineKills, vaccinekills, StopTheVaccines, fuckyourvaccine, VaccineIssues, vaccinesDONTwork, vaccinebad, VaccineNotTheAnswer, dontgetthecovidvaccine, MurderbyVaccine, vaccineforwhat