Can We Stop Fake News?
Using Agent-Based Modelling to Evaluate Countermeasures for Misinformation on Social Media

Anna Gausen, Wayne Luk and Ce Guo
Department of Computing, Imperial College London, UK
{anna.gausen16, w.luk, c.guo}@imperial.ac.uk

Abstract

The internet and social media have been a huge force for change in our society, transforming the way we communicate and seek information. However, the unprecedented volumes of data have created a unique set of challenges: misinformation, polarization and online conflict. Misinformation can be particularly detrimental to society, democracy and public health. As a result, there is a growing area of research into countermeasures against online misinformation. However, much of this research is conducted in small scale experiments and cannot predict the macro-level impact. This paper demonstrates that agent-based modelling can be a useful tool for policy-makers to evaluate these countermeasures at scale before implementing them on a social media platform. This research has the following contributions: (i) Development of an agent-based model of the spread of information on a social media network, based on Twitter. (ii) Calibration and validation of the proposed model by Twitter data following fake and true news stories. (iii) Using agent-based modelling to evaluate the impact of countermeasures on the spread of fake news and on general information sharing.

Introduction

Social media has transformed the way we get informed, share ideas and debate. The unprecedented scale of these platforms and the absence of moderation have led to a misinformation "infodemic" (Naeem, Bhatti, and Khan 2020). This has impacted presidential campaigns (Howard et al. 2017), spread hate (Woolley and Howard 2016) and propagated fake news (Lazer et al. 2018).

Across the industry, there is significant research into mitigating the impact of misinformation online. Due to the complex nature of the propagation of information, it can be challenging to predict the impact of countermeasures before they are implemented on a live social media platform. Behavioural countermeasures are a new approach that have been effective in behavioural experiments (Lewandowsky and van der Linden 2021) (Pennycook et al. 2021). However, these experiments cannot account for the macro-scale impact and emergent behaviour when implemented on a network.

This paper presents an agent-based model of a social media network. The model is based on Twitter, where the agents, who represent social media users, can share information with their neighbours or followers. This is inspired by epidemic modelling and has been seen in related research. Our approach addresses two challenges which have not been satisfactorily covered in existing models: the limited use of real datasets for validation, and the inability to evaluate countermeasures.

Our approach has two contributions. The first is an agent-based model capable of evaluating the impact of misinformation countermeasures on the spread of both true and fake news stories on social media. The second is the use of news stories related to the current COVID-19 pandemic to calibrate and validate the model. These contributions will help tackle the challenge of online misinformation, by enabling decision-makers to evaluate countermeasure performance on a model before applying them to a real population.

Background and Related Work

Agent-Based Modelling and Social Media

There has been a long tradition of modelling the spread of ideas through social networks (Kendall and Daley 1964). Agent-based modelling (ABM) "consists of describing a system from the perspective of its constituent units" (Bonabeau 2002). These have simplistic behaviours at an agent-level which can combine to produce unexpected results. ABMs are appropriate for representing social media networks as they involve interactions between high numbers of heterogeneous agents and exhibit emergent behaviour (Kyamakya 2006).

Table 1 reviews how related work addresses the issues below. Does the ABM: (1) Model the spread of misinformation? (2) Model a social media network? (3) Use real data for calibration and empirical validation? (4) Evaluate the effectiveness of countermeasures for misinformation?

The review highlights challenges not addressed by existing ABM research. Firstly, there is limited use of real social media data and, secondly, most current efforts do not cover evaluation of countermeasures. An exception is (Sathanur, Sui, and Jandhyala 2015), which simulates strategies for controlling rumors but, without validation or grounding in real data, it cannot provide a convincing evaluation of their impact. Based on this review, the proposed model is novel in its ability to address all four issues in Table 1.

Table 1: Review of Existing ABM Research

Issues:	1	2	3	4
Fan, Xu, and Zhao (2018)		Y		
Alvarez-Galvez (2016)	Y			
Sathanur, Sui, and Jandhyala (2015)	Y	Y		Y
Kaligotla, Yucesan, and Chick (2016)	Y	Y	Y	
Jin et al. (2013)	Y	Y	Y	
Tang et al. (2014)		Y		
Onuchowska and Berndt (2020)	Y	Y		
Serrano and Iglesias (2016)	Y	Y	Y	
Our Approach	Y	Y	Y	Y

Countermeasures for Misinformation

Designing countermeasures for misinformation is a growing area of research. Commonly, the focus is on fact-checking information. This approach, whilst important, relies on accuracy in detection, often at the detriment of speed. This paper focuses on alternative interventions: rule-based policies, societal inoculation and accuracy prompting.

Rule-Based Policies are enforced by social media platforms to limit the spread of misinformation online through blocking users and removing posts based on complaints.

Societal Inoculation protects individuals by building "mental antibodies" to future misinformation (Goldberg 2021). Inoculation warnings and pre-bunking help users resist the "sticky" nature of misinformation (Lewandowsky et al. 2020).

Accuracy Flags could be a scalable method to mitigate online misinformation. A recent study has showed that distraction leads to increased engagement with misinformation (Pennycook et al. 2021). By shifting user's attention to the concept of accuracy can improve the quality of information that they share.

Proposed Model Description

The proposed model is based on the ABM developed in (Serrano and Iglesias 2016). It has three extensions: (i) The proposed model introduces the concept of an influential agent, since the influence or perceived credibility of the source is an important factor as to whether a person re-shares information (Onuchowska and Berndt 2020). (ii) It provides the ability to implement countermeasures on the network and evaluate their effectiveness. (iii) It is empirically validated with both true and fake news stories related to COVID-19.

Proposed Model without Countermeasures

Initially, the proposed model is run without countermeasures for validation with real data and comparison to the reference model (Serrano and Iglesias 2016).

The proposed model has four agent states: (i) Susceptible: users who are susceptible to (mis)information, this will be the majority of the population at the start; (ii) Believe: users that believe and try to convince other users of the (mis)information, also known as infecting them; (iii) Deny: users that do not believe the (mis)information and try to convince other users that it is false, also known as vaccinating them; (iv) Cured: users that previously believed the (mis)information but no longer do. These users stop sharing information on the subject.

The model behaviour can be described qualitatively as: (i) Initialise the number of infected agents (State=Believe). (ii) At each timestep the infected users try to infect their susceptible neighbours with a probability of P_{inf}. However these neighbours may become deniers with a probability of P_{deny}. (iii) Vaccinated users (State=Deny) try to vaccinate their susceptible neighbours with a probability of P_{vacc}. (iv) Vaccinated users try to cure their infected neighbours with probability of P_{deny}. (v) An influential user's ability to infect or vaccinate will increase by P_{infl}.

Proposed Model with Countermeasures

Once the proposed model without countermeasures is calibrated and empirically validated, the model can be simulated with countermeasures to evaluate their impact.

This research explores the effect of the rule-based policy of user blocking. If a user receives a number of complaints their account is blocked, preventing them from interacting with other users. This is implemented by introducing an additional state $State.Blocked$, an agent attribute (N_{comp}) denoting the number of complaints, and the probability that a vaccinated user decides to file a complaint against a user sharing (mis)information P_{block}. For this study, N_{comp} = 3 and P_{block} = 0.1.

Societal inoculation describes the process of providing warnings about misinformation to increase an individual's resistance to it. In terms of the model, this means that a user will be less likely to become infected:

$$P_{inf} := P_{inf} - P_{inoc} \tag{1}$$

where $:=$ denotes updating the value of the probability and P_{inoc} is the experimental impact of general inoculation warnings. The value used in this model is calibrated using the results published in a recent paper (Lewandowsky and van der Linden 2021) based on (van der Linden et al. 2017). This found general inoculation policies cause a 6.5% reduction in the likelihood of believing misinformation.

Accuracy prompts increase the likelihood that a user correctly assesses whether a piece of information is accurate:

$$P_{vacc} := P_{vacc} + P_{acc} \tag{2}$$

where P_{acc} is the calculated experimental impact of accuracy prompts. This value is calibrated from an experiment that found, on average, accuracy flags improved the quality of posts shared by 4.8% (Pennycook et al. 2021).

Social Media Dataset

The CoAID dataset is used to calibrate and empirically validate the ABM (Cui and Lee 2020). It shows Twitter user-engagement with fake and real news stories about the current COVID-19 epidemic. It includes ground truth labels for the news stories and 296,752 user engagements from between 1st May 2020 to 1st November 2020. These data include Tweet ID, News ID and replies for labelled news stories. For

this research, we use the Twitter Developer API and TweePy Python toolbox to generate the timestamp information for each Tweet ID in the CoAID dataset. This information is required to calibrate and validate the proposed model.

Evaluation

Experimental Set-up

Three experiments are carried out: (1) Validation: run proposed model with no countermeasures and compare to output of S&I model (Serrano and Iglesias 2016). (2) Fake: run proposed model with and without countermeasures for CoAID fake news story. (3) True: run proposed model with and without countermeasures for CoAID true news story.

Table 2: Table with configured and tuned parameters for each experiment.

	Runs	**Validation**	**Fake**	**True**
Config	**Timesteps**	55	134	60
Params	**Init. Inf %**	4	4	0.5
	Prob. Inf.	0.02	0.01	0.0035
Tuned	**Prob. Vacc.**	0.02	0.1	0.02
Params	**Prob. Deny**	0	0.002	0.015
	Prob. Infl.	0.05	0.005	0.06

Three experiments are carried out with 10,000 agents for 10 simulations. Table 2 shows the configured and tuned parameters for each experiment. The tuned parameters are optimised by minimising the calculated root mean square error (RMSE) between the simulation output of the proposed model with no countermeasures and the real data-points. This process is repeated until the optimal probability values are found. Then the normalised RMSE of the simulation output is given by:

$$NRMSE = \frac{RMSE}{P_{max} - P_{min}} \tag{3}$$

where P is the proportion of users sharing misinformation in the real dataset. This enables model validation and comparisons between experiments (Kell, Forshaw, and McGough 2020). For this research, provided $NRMSE < 0.2$ the model is seen as valid.

Once the proposed model with no countermeasures has been empirically validated, the countermeasures can be applied to the network. The three countermeasures are implemented independently and their parameters are calibrated from real experimental results. The following two metrics are used to evaluate the impact of each countermeasure: (1) The percentage change in the maximum proportion of the population sharing the (mis)information (ΔP_{max}). (2) The percentage change in the average proportion of the population sharing the (mis)information (ΔP_{avg}). Both given by:

$$\Delta P(\%) = 100 \times \frac{P^{cm} - P^{pm}}{P^{pm}} \tag{4}$$

where pm represents the proposed model with no countermeasures, cm represents the proposed model with a countermeasure, P is the proportion of the population sharing the (mis)information.

Model Validation

The model is validated by comparing the output of the proposed model with no countermeasures to the S&I model. This is a reference model replicating the ABM described in (Serrano and Iglesias 2016). The comparison is based on the dataset in (Serrano and Iglesias 2016), which contains 4,423 tweets about the rumour that Sarah Palin is divorced. This is used to validate that the additional logic in the proposed model provides a more realistic description of the information spreading dynamics on a social media network.

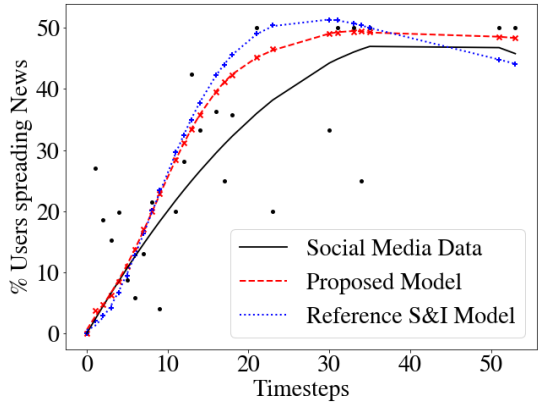


Figure 1: Graph comparing the output of the proposed model with no countermeasures (red), the output of S&I model (blue) and Palin dataset (black). For the proposed model: RMSE is 8.70 and NRMSE is 0.17. For the S&I model: RMSE is 9.80 and NRSME 0.20.

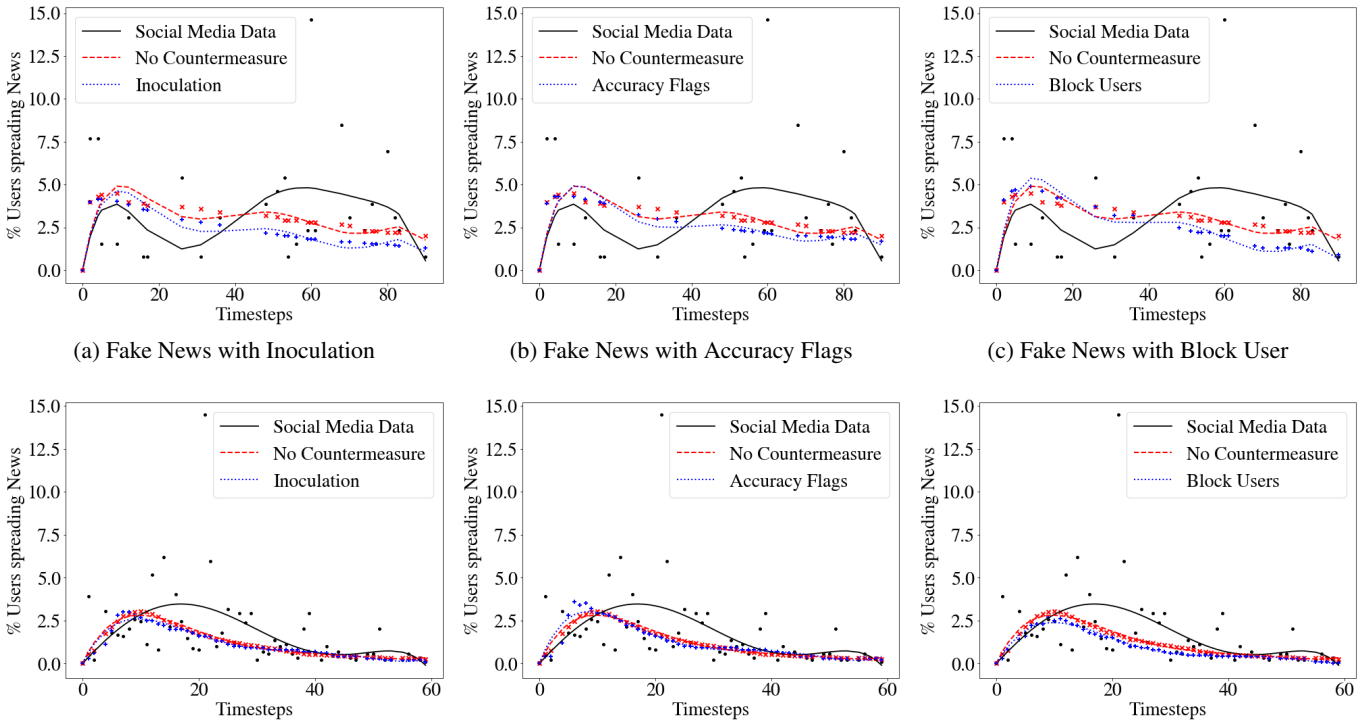
Figure 1 shows the comparison between the proposed model and the reference S&I model. For a fair comparison, the probabilities in both models are tuned to minimise the RMSE from the real data-points. This analysis shows that the simulation output of the proposed model better replicates the spread of the Palin rumour than the reference S&I model.

Evaluation of Countermeasures

This section evaluates the impact of the countermeasures on the spread of both real and fake news stories. It is important to note that agent-based models provide useful insights for the comparison of policies; they do not however provide exact predictions. This is due to the in-built assumptions and simplifications.

Figure 2 presents the average impact of each countermeasure on the propagation of both a fake and a true COVID-19 news story (Cui and Lee 2020), over 10 simulations. Table 3 shows the calculations of the countermeasure evaluation metrics for both news stories. The results are shown as percentage differences in the proportion of users who believe the information compared to the proposed model with no countermeasures.

The results indicate that all the countermeasures would be effective in reducing the spread of misinformation. Based on these results, inoculation is shown to be the most effective. However, its performance is very similar to blocking users, making it difficult to confidently evaluate which



(a) Fake News with Inoculation (b) Fake News with Accuracy Flags (c) Fake News with Block User

(d) True News with Inoculation (e) True News with Accuracy Flags (f) True News with Block User

Figure 2: Figures (a)-(c) show the spread of fake news story: "China stole Coronavirus from Canada and weaponized it into a Bioweapon". For model with no countermeasures: RMSE is 2.27 and NRMSE is 0.16. Figures (d)-(f) show the spread of true news story: "What is COVID-19?". For model with no countermeasures: RMSE is 1.06 and NRMSE is 0.07.

Table 3: Evaluation of Misinformation Countermeasures

	Run	$\Delta P_{avg}(\%)$	$\Delta P_{max}(\%)$
Fake News	Inoculation	-35.87	-7.78
	Accuracy Flags	-25.68	-2.44
	Block Users	-33.01	11.11
True News	Inoculation	-12.10	-1.96
	Accuracy Flags	-4.84	17.65
	Block Users	-24.73	-11.76

would be most effective. The countermeasure impact on the true news story provides additional insight. It shows that the behavioural countermeasures have much less impact on the propagation of the true news story than blocking users.

This indicates that the behavioural countermeasures, inoculation and accuracy flags, would be the most effective at minimising the spread of misinformation whilst having the least impact on the spread of truthful information. This is an important balance for social media platforms to achieve as their functionality should not be hindered by misinformation countermeasures.

Conclusion

The phenomenon of online misinformation poses one of the biggest challenges of the coming decade. Mitigating the impact of misinformation on social media is essential to foster democratic debate, strengthen civic discourse and promote access to trustworthy news. This research shows that agent-based models of social media networks are a useful tool for evaluating countermeasures for misinformation. These preliminary results indicate that they could allow policy-makers to test the impact of proposed countermeasures, on both true and fake news, before implementing them on a live social media platform.

This research represents an initial step towards using agent-based modelling in the fight against misinformation. The proposed model has a number of limitations such as a non-realistic population size, a lack of formal verification and a limited number of datasets.

Future work in this field could help realise the full potential for a scalable and verifiable tool to evaluate policies to mitigate misinformation online. Three directions that should be considered are formal verification of the model output, enhanced capability of our approach to capture social media user behaviour, and computational acceleration to model more realistic population sizes and behaviours.

Acknowledgements

The support of UK EPSRC (grant number EP/S023356/1, EP/L016796/1 and EP/P010040/1), Intel and Xilinx is gratefully acknowledged.

References

Alvarez-Galvez, J. 2016. Network Models of Minority Opinion Spreading: Using Agent-Based Modeling to Study Possible Scenarios of Social Contagion. *Social Science Computer Review* 34(5): 567–581.

Bonabeau, E. 2002. Agent-Based Modeling: Methods and Techniques for Simulating Human Systems. *Proceedings of the National Academy of Sciences of the United States of America* 99(SUPPL. 3): 7280–7287.

Cui, L.; and Lee, D. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. URL https://github.com/cuilimeng/CoAID.

Fan, R.; Xu, K.; and Zhao, J. 2018. An Agent-Based Model for Emotion Contagion and Competition in Online Social Media. *Physica A: Statistical Mechanics and its Applications* 495: 245–259.

Goldberg, B. 2021. Can "Inoculation" Build Broad-Based Resistance to Misinformation? URL https://medium.com/jigsaw/can-inoculation-build-broad-based-resistance-to-misinformation-6c67e517e314.

Howard, P. N.; Bradshaw, S.; Kollanyi, B.; Desigaud, C.; and Bolsover, G. 2017. Junk News and Bots during the French Presidential Election: What Are French Voters Sharing Over Twitter? *Comprop Data Memo* 5(May): 1–5.

Jin, F.; Dougherty, E.; Saraf, P.; Cao, Y.; and Ramakrishnan, N. 2013. Epidemiological Modeling of News and Rumors on Twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, 1–9.

Kaligotla, C.; Yucesan, E.; and Chick, S. E. 2016. An Agent Based Model of Spread of Competing Rumors through online interactions on social media. *Proceedings - Winter Simulation Conference* 2016-Febru: 3088–3089.

Kell, A. J.; Forshaw, M.; and McGough, A. S. 2020. Long-Term Electricity Market Agent Based Model Validation Using Genetic Algorithm Based Optimization. In *e-Energy 2020 - Proceedings of the 11th ACM International Conference on Future Energy Systems*, 1–13. Association for Computing Machinery, Inc. ISBN 9781450380096.

Kendall, D. J.; and Daley, D. G. 1964. Epidemics and Rumours. *Nature, 204(4963):1118–1118* .

Kyamakya, K. 2006. Artificial Intelligence in Transportation Telematics. *OGAI Journal (Oesterreichische Gesellschaft fuer Artificial Intelligence)* 25(3): 2–4.

Lazer, D. M. J.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; Schudson, M.; Sloman, S. A.; Sunstein, C. R.; Thorson, E. A.; Watts, D. J.; and Zittrain, J. L. 2018. The Science of Fake News. *Science* 359(6380): 1094–1096.

Lewandowsky, S.; Cook, J.; Ecker, U. K. H.; Albarracín, D.; Amazeen, M. A.; Kendeou, P.; Lombardi, D.; Newman, E. J.; Pennycook, G.; Porter, E.; Rand, D. G.; Rapp, D. N.; Reifler, J.; Roozenbeek, J.; Schmid, P.; Seifert, C. M.; Sinatra, G. M.; Swire-Thompson, B.; van der Linden, S.; Vraga, E. K.; Wood, T. J.; and Zaragoza, M. S. 2020. Debunking Handbook 2020. URL https://sks.to/db2020.

Lewandowsky, S.; and van der Linden, S. 2021. Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology* 1–39.

Naeem, S. B.; Bhatti, R.; and Khan, A. 2020. An Exploration of How Fake News is Taking Over Social Media and Putting Public Health at Risk. *Health Information and Libraries Journal* 1–7.

Onuchowska, A.; and Berndt, D. J. 2020. Using Agent-Based Modelling to Address Malicious Behavior on Social Media. In *Proceedings of the 40th International Conference on Information Systems, ICIS 2019*. ISBN 9780996683197.

Pennycook, G.; Epstein, Z.; Mosleh, M.; Arechar, A. A.; Eckles, D.; and Rand, D. G. 2021. Shifting Attention to Accuracy Can Reduce Misinformation Online. *Nature* 1–6.

Sathanur, A. V.; Sui, M.; and Jandhyala, V. 2015. Assessing Strategies for Controlling Viral Rumor Propagation on Social Media - A Simulation Approach. *2015 IEEE International Symposium on Technologies for Homeland Security, HST 2015* 1–6.

Serrano, E.; and Iglesias, C. A. 2016. Validating Viral Marketing Strategies in Twitter via Agent-Based Social Simulation. *Expert Systems with Applications* 50: 140–150.

Tang, M.; Mao, X.; Yang, S.; and Zhu, H. 2014. Policy Evaluation and Analysis of Choosing Whom to Tweet Information on Social Media. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 45–50. IEEE.

van der Linden, S.; Leiserowitz, A.; Rosenthal, S.; and Maibach, E. 2017. Inoculating the Public against Misinformation about Climate Change. *Global Challenges* 1(2): 1600008.

Woolley, S. C.; and Howard, P. N. 2016. Political Communication, Computational Propaganda and Autonomous Agents Introduction. *International Journal of Communication 10(2016)* 4882–4890.