How Gullible Are You? Predicting Susceptibility to Fake News

Tracy Jia Shen
jqs5443@psu.edu
The Pennsylvania State University
University Park, PA, USA

Robert Cowell
rhc5082@psu.edu
The Pennsylvania State University
University Park, PA, USA

Aditi Gupta
ajg6035@psu.edu
The Pennsylvania State University
University Park, PA, USA

Thai Le
tql3@psu.edu
The Pennsylvania State University
University Park, PA, USA

Amulya Yadav
amulya@psu.edu
The Pennsylvania State University
University Park, PA, USA

Dongwon Lee
dongwon@psu.edu
The Pennsylvania State University
University Park, PA, USA

ABSTRACT

In this research, we hypothesize that some social users are more *gullible* to fake news than others, and accordingly investigate on the susceptibility of users to fake news–i.e., how to identify susceptible users, what are their characteristics, and if one can build a prediction model.Building on the crowdsourced annotations of 5 types of susceptible users in Twitter, we found out that: (1) susceptible users are correlated with a combination of user, network, and content features; (2) one can build a reasonably accurate prediction model with 0.82 in AUC-ROC for the multinomial classification task; and (3) there exists a correlation between the dominant susceptibility level of center nodes and that of the entire network.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Machine Learning** → *Social Media.*

KEYWORDS

Fake news, user susceptibility, machine learning

ACM Reference Format:
Tracy Jia Shen, Robert Cowell, Aditi Gupta, Thai Le, Amulya Yadav, and Dongwon Lee. 2019. How Gullible Are You? Predicting Susceptibility to Fake News. In *11th ACM Conference on Web Science (WebSci '19), June 30–July 3, 2019, Boston, MA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292522.3326055

1 INTRODUCTION

Since the 2016 US election, the public interest on fake news has exploded, resulting in many early solutions to computationally detect fake news and understand how it spreads in social media. While they are effective solutions to the problem of fake news, however, we believe that another important problem that has received little attention to date is to study on users' susceptibility to fake news,

Class	Tweets
Strong Agreement	Always knew these kids had ties to #Soros . The parents need to have a Mental background check.
Weak Agreement	Florida is his home base! He found more Useful Idiots!
Neutral	They all look SO saddened by the loss of their classmates....don't they???!!!
Weak Disagreement	Youre a con man too.
Strong Disagreement	Wow. Imagine being a bad enough human to attack teenagers who just survived a mass shooting.

Figure 1: Five levels of susceptibility and their examples.

i.e., why and how some people become susceptible to fake news (while others do not). In this work, therefore, we ask two research questions (RQs): (1) what are the characteristics of Twitter users who are susceptible (or not susceptible) to fake news?, and (2) is it possible to build an accurate model to predict susceptible users in Twitter? To answer these RQs, first, we need to create labeled samples (i.e., Twitter users who are determined to be susceptible to fake news). As there exists no such dataset, however, we instead resort to crowdsourced samples with an assumption that Twitter users who express their agreement or disagreement toward verified fake news are treated as *susceptible* or *not-susceptible* users. Figure 1 shows real Tweet examples of five degrees of agreement to fake news–highly-susceptible (= strong agreement), slightly-susceptible (= weak agreement), neutral, not-quite susceptible (= weak disagreement), and not-at-all susceptible (= strong disagreement).

There exists very few prior work which studies user's susceptibility to fake news. For instance, [1] studied user susceptibility as a behavioral factor to predict viral diffusion of general information rather than fake news. [5] studied user susceptibility to fake news created by bot activities and limit their definition of 'susceptible users' as the ones that interacted at least once with a social bot. [3] studied research questions similar to ours, but used the methodology from cognitive science. Unlike these works, we focus on multiple degrees of susceptibility to fake news by employing machine learning methods.

2 EXPERIMENT AND PREDICTION

Dataset: We first identified 7 verified fake news about the "Parkland Shooting" incident where a gunman killed 17 students and staffs at Marjory Stoneman Douglas High School in Parkland, Florida on Feb. 14, 2018. We then identified 896 users who replied to the 7 fake news from Feb 14 to May 20, 2018, and scraped additional 13,000

WebSci '19, June 30-July 3, 2019, Boston, MA, USA

ACM ISBN 978-1-4503-6202-3/19/06.
https://doi.org/10.1145/3292522.3326055

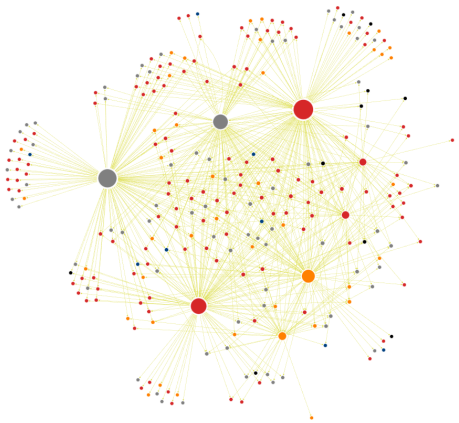

Figure 2: Center nodes network (the biggest 8 nodes), where red=strong-agreement, orange=weak-agreement, grey=neutral, blue=weak-disagreement, and black=strong-disagreement.

Twitter users who were followers or friends of the repliers). Next, for each reply, we assign 5 English-literate Amazon Mechanical Turk (MTurk) workers (≥ 95% approval rate) to label it into one of five classes. We used the majority voting to determine the final class for each reply.

Preliminary Analysis: We checked the spreading network to see if susceptibility is infectious or not. Since we don't have labels for the followers/friends, we only formed the repliers' susceptibility network. The largest component in the network contained 63% of all nodes, including 8 center nodes, 5 of which have more than 100 connections (following relationships). Among 8 center nodes, 4 were labeled as strong agreement (bigger red nodes in Figure 2). Among the nodes connected to these 8 center nodes, about 51% were strong agreement, when strong agreement users accounted for 37% of the entire network. Therefore, it appears that the dominant susceptibility level of center nodes in the largest component of a network is correlated with the dominant susceptibility of the entire network, if majority of center nodes are labeled as strong agreement (in our case, 50%).

Feature Engineering: We considered three types of features: (1) *Content:* linguistic (LIWC) features including usage of punctuation, latent emotions, perception, cognitive thinking in Twitter posts; (2) *User:* # followers, # friends, # lists, # statuses, create_time, circulation_time, and membership_year; and (3) *Network:* clustering features (measuring how users cluster), centrality features (measuring how close users are), influence and special connection features.

Prediction Result: We used four learning models (e.g., KNN, Random Forest, Decision Tree, and XGBoost) to solve the proposed multinomial classification task. Due to the scarcity of training samples, we did not test data-hungry algorithms such as deep learning algorithms. We evaluated the learned models using the Area-Under-the-Curve of the Receiver Operating Characteristics (AUC-ROC) to judge the robustness of the model [2, 4]. Using four learning models, we compared the prediction performance of seven combinations of features–i.e., content-only, user-only, network-only, content + user, content + network, user + network, and all features combined. The experimental results are summarized in Table 1. In Figure 3, using the best performing XGBoost model to predict both strong-agreement and strong-disagreement data (i.e., highly-susceptible and not-at-all susceptible users), we have over 0.83 AUC performance. The model, however, does not predict the minority classes well, with only 0.54 for weak-agreement and 0.69 for weak-disagreement in AUC. However, the poor performance for minority classes could be due to the scarcity of training data.

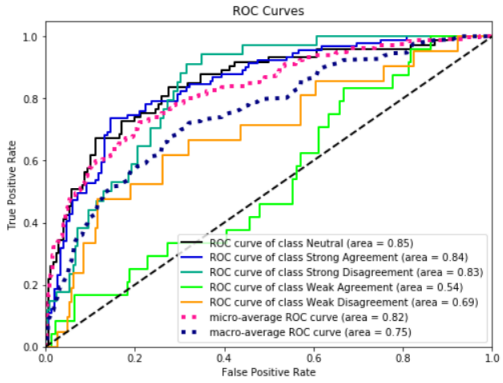

Figure 3: ROC curves of all five classes by XGBoost.

Table 1: Micro-AVG AUC-ROC for all classes

	Content	User	Network	Content + User	Content + Network	User + Network	All Features
KNN	0.60	0.59	0.63	0.65	0.68	0.61	0.62
Decision Tree	0.54	0.62	0.56	0.68	0.57	0.64	0.65
Random Forest	0.66	0.74	0.68	0.71	0.68	0.73	0.77
XGBoost	0.69	0.76	0.68	0.8	0.75	0.77	0.82

3 CONCLUSION

First, we found that the high susceptibility level of center nodes has high correlation with the entire network susceptibility level. Second, we demonstrated that it is possible to differentiate one of five susceptibility levels of users using various features trained in XGBoost model, achieving 0.82 in AUC-ROC.

4 ACKNOWLEDGEMENT

This work was in part supported by NSF awards #1742702 and #1820609, and ORAU-directed R&D program 2018.

REFERENCES

[1] Tuan-anh Hoang and Ee-peng Lim. 2012. Virality and Susceptibility in Information Diffusions. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* 2010 (2012), 146–153.

[2] M Hossin and Sulaiman. 2015. A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS. *International Journal of Data Mining & Knowledge Management Process (IJDKP)* 5, 2 (2015). https://doi.org/10.5121/ijdkp.2015.5201

[3] Gordon Pennycook and David G. Rand. 2018. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* September 2017 (2018), 1–12. https://doi.org/10.1016/j.cognition.2018.06.011

[4] Sebastian Raschka. [n. d.]. Machine Learning FAQ. https://sebastianraschka.com/faq/docs/multiclass-metric.html

[5] Claudia Wagner, Markus Strohmaier, Silvia Mitter, and Christian Körner. 2012. When social bots attack : Modeling susceptibility of users in online social networks. *#MSM2012 Workshop proceedings* (2012), 41–48. http://ceur-ws.org/Vol-838