Trust Models and Social Networks: How to Slow the Spread of Online Misinformation

Casey Bates
Santa Clara University
Santa Clara, California, United States of America
cbates@scu.edu

Abstract

As more people come to rely on online social networks as a source for information more so than traditional news sources, these platforms become communication platforms where anyone can share anything they wish. Along with this quick and easy exchange of information comes the risk of a rapid spread of misinformation. In an attempt to stop the spread of potentially harmful misinformation across their platforms, we demonstrate how online social networks such as Facebook can implement a user based trust model such to detect the likelihood of a user posting misinformation information, and stop them from actively spreading it. Using real Facebook post fact checked by Buzzfeed journalists, our results show that this method can be an effective tool to stop misinformation from spreading across social media.

***Index Terms*—Misinformation, Trust Model, Beta Reputation Model, Online Social Network**

I. Introduction

The world we live in is dominated by technology. Seemingly every industry and every way of life are influenced in part by technology and the rapid spread of information it has caused. Unfortunately, not all of this information is factual. The spread of misinformation is at an all-time high. This proliferation of misinformation is occurring at an alarming rate, and it is extremely damaging to our society. As seen with recent presidential elections, the spread of misinformation has helped to cause a political divide in the United States. There have even been cases of foreign states using misinformation to attempt to influence recent elections [1]. Public health emergencies like COVID-19 have been exacerbated by the spread of misinformation as well. Many have been misled about the effects and treatments of COVID-19, causing some to distrust the scientific community's safety precautions and partake in potentially deadly alternate treatments [2]. Through online social networks such as Facebook, Twitter, and YouTube, this kind of misinformation can spread like wildfire, with little to slow it.

II. Related Work and Datasets

It has become increasingly necessary to be able to detect and stop the spread of this type of information, but these platforms have seen little success. The identification of misinformation can be a slow and manual process, as it can take time for a human to fact-check. In addition, misinformation often spreads rapidly due to their often shocking or interesting nature, leading to a much faster spread than truthful information and not leaving much time for fact-checking. Most recently, this has occurred with misinformation regarding the COVID-19 pandemic. In an effort to combat this potentially deadly public health misinformation and improve existing misinformation detection methods, Raju et al. developed a detection model using transfer learning to identify and remove COVID-19 misinformation on Twitter [3]. Similarly, Buntain and Golbeck present a model for automatic detection of "fake news" using feature analysis of Twitter threads [4].

Analyzing the content of posts themselves is not the only way to detect misinformation. Xu et al. demonstrates a method of misinformation detection by computing the domain reputation of the publisher, using statistics such as user registration data of the publisher's site, site age, domain age, domain popularity, and probability of news disappearance [5]. This research also found that content understanding using a combination of TF-IDF (term frequency - inverse document frequency) analysis and term and query vector similarity, can also be an effective way to detect possible misinformation in social networks [5]. Yet another solution to this issue is the use of user trust modeling to determine the probability of misinformation spread by individual users. Bodnar et al. looks at a user's physical distance and time from an event to determine the likelihood that they believe it actually took place [6].

It is also the case that not all misinformation being spread across a network will not always be observed. Zhang et al. present a possible solution, strategically placing monitors throughout the network in order to identify the diffusion of misinformation [7]. In addition to placing monitors to detect the spread of misinformation, it may also be important to identify where the misinformation originated from. Maryam and Ali discuss a depth first search source identification model that can be used to traverse a network and identify the origin of diffused misinformation [8].

Although platforms like Facebook and Twitter have attempted to implement their own detection systems for misinformation, that have not always been adequate, and users may not always have the resources to verify information themselves. For example, Facebook's system relies on a user reporting system, where users flag posts for third-party fact checkers to investigate [9]. Pourghomi et al. propose an alternative approach called “right-click authenticate” where users have access to a fact checking tool in the right-click context menu in their web browsers [9]. This solution would give all users the ability to verify information themselves, without relying on platforms like Facebook to combat misinformation for them.

A. BuzzFeed News Fact-Check Dataset

In this research, we used the Fact-Checking Facebook Politics Pages dataset collected by BuzzFeed News [10] as the source of information. This dataset includes 2282 Facebook posts and associated engagement statistics (as of October 11, 2016) such as share count, reaction count, and comment count, from nine unique users. BuzzFeed journalists fact-checked each post and assigned it a rating based on its content: ”mostly false”, ”mixture of true and false”, ”mostly true”, and ”no factual content”.

III. Methods

While most existing research in this area focuses mainly on identifying misinformation amongst posts being shared [3] [4] [5] [6] or where the false information originated from [8], we decided to look at how to slow the spread of misinformation once it has been posted.

To attempt to detect which users in the network are at risk of spreading misinformation, we will employ variations of a Beta Reputation Trust model to assign trust values to each user. Using the trust values calculated by each model, users with low reputations will be penalized by simulating hiding posts once their trust value falls below a certain threshold value. The hidden posts will not be seen by other users in the network, and thus will not have any engagement, preventing the post from being shared further. Ideally, this will provide some insight into how the trust models might be used to lessen the number of people impacted by the misinformation.

A. Beta Reputation Model

Where other misinformation detection systems solely remove posts marked as misinformation [3], the beta reputation model will also keep trust scores for each user, making it trivial to determine which users are most likely to spread false information.

Based on past research, many information classification models are not able to label a post as strictly “true” or “false” [3] [4], which can occur for a few reasons. For one, existing machine learning or other classification models can only give a likelihood of a post being misinformation with a certain amount of confidence. Another is that the content itself is often not entirely true or entirely false. An issue arises when considering a prevention model that removed any posts labeled as containing misinformation, as this type of model would rely on a binary system where posts are either misinformation or not. The beta reputation model, however, is capable of handling a range of post ratings. For example, post ratings could range from 0 to 1, where a post with a value of 0.8 represents 80% misinformation and 20% truthful information, or posts could be assigned one of “Mostly True”, “Mixture of True and False”, and “Mostly False”. In both scenarios, the beta reputation model would be able to calculate a trust value that represents the likelihood that a user will post truthful information.

However, the beta reputation model is not foolproof, as a user’s trustworthiness is not a perfect predictor of what type of content they will post. A predominantly trustworthy user’s next post could be misinformation, in which case no action would be taken, as their trust value might still be higher than the threshold to remove their post, leading to misinformation slipping through the cracks and spreading across the network. At the same time, a historically untrustworthy user might upload a post that is completely factual, yet due to their trust value being so low, the post with true information will still be taken down.

To compare iterations of the beta reputation model, we used the share count value of each post, each user, and each category of post rating. Using the share count as our comparison statistic gives us an indicator of how often information is being spread across the network. We define a successful model as one that has fewer shares of posts rated “mostly false” than if no model were implemented, as this would indicate less misinformation is being seen and disseminated across the network.

B. Calculating Trust

To determine trust values for each user, posts were iterated through in chronological order, simulating each user “posting” their content. For each post, the BuzzFeed journalist’s ratings were mapped to an integer between 0 and 2, as seen in Table 1. Posts rated “no factual content” did not factor into the calculation, as they contained only advertisements, opinion pieces, or other content not claiming to be factual.

TABLE I
Post Rating Mappings

Post Rating	Integer Value
mostly false	0
mixture of true and false	1
mostly true	2

A user’s trust value (1) was then calculated immediately after their post’s rating was determined. Let V be the set of previous trust values.

$$T = \frac{(2 * R_t^x) + (0 * S_t^x) + 2}{R_t^x + S_t^x + 2} \tag{1}$$

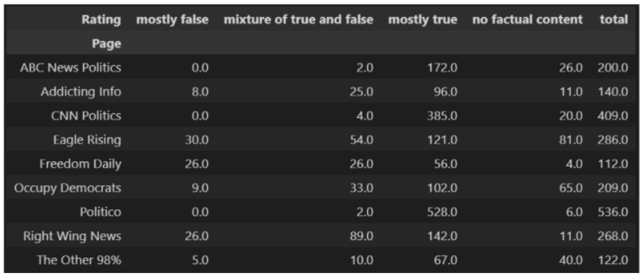
where,

$$R_t^x = \sum_{r \in V} \frac{r - 0}{2 - 0} \tag{2}$$

$$S_t^x = \sum_{r \in V} 1 - \frac{r - 0}{2 - 0} \tag{3}$$

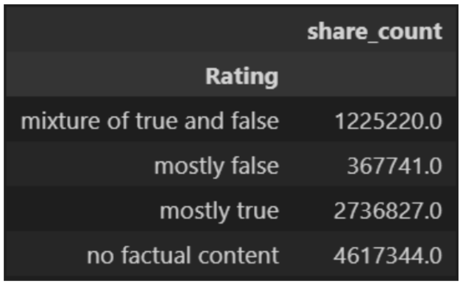
IV. Results

Iteration zero of this model was used as a control, where no posts were removed. This iteration did not make any modifications to the original dataset, it only counted the number of posts in each rating category for each page, and the share counts of all posts in each rating category. As seen in 1, there were 104 total posts rated “mostly false” by BuzzFeed’s journalists, as well as 245 posts rated “mixture of true and false”. As illustrated by 2, there were 367,741 shares of “mostly false” posts and 12,25,220 shares of posts containing a “mixture of true and false”. Without making any changes to how misinformation was handled on Facebook at the time the data was collected, posts containing at least some misinformation were shared 15,92,961 times.



Rating	mostly false	mixture of true and false	mostly true	no factual content	total
Page					
ABC News Politics	0.0	2.0	172.0	26.0	200.0
Addicting Info	8.0	25.0	96.0	11.0	140.0
CNN Politics	0.0	4.0	385.0	20.0	409.0
Eagle Rising	30.0	54.0	121.0	81.0	286.0
Freedom Daily	26.0	26.0	56.0	4.0	112.0
Occupy Democrats	9.0	33.0	102.0	65.0	209.0
Politico	0.0	2.0	528.0	6.0	536.0
Right Wing News	26.0	89.0	142.0	11.0	268.0
The Other 98%	5.0	10.0	67.0	40.0	122.0

Fig. 1. Raw Count of Fact-Checked Posts for Each User



	share_count
Rating	
mixture of true and false	1225220.0
mostly false	367741.0
mostly true	2736827.0
no factual content	4617344.0

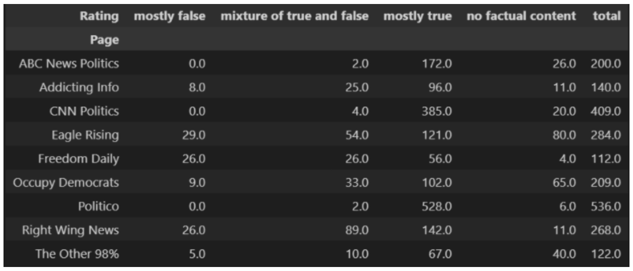
Fig. 2. Raw Count of Posts in Each Rating Category

A. Post Removal Below 50% Trust

The first iteration of the model implemented a trust value calculation for each user after each post, however, if their trust value fell below 50% of the maximum trust, the post would be removed. As shown in Figure 3, this iteration did not have much of an effect on the network. While one “mostly false” post was removed, there were still 245 posts rated “mixture of true and false” this version of the model, and one post labeled “no factual content” was removed. Figure 4 paints a similar picture, as the share count of “mostly false” posts only decreased by 7,577. With only 2 posts being removed and only 2.06% of shares of misinformation being prevented, this iteration of the model was not as effective as expected.

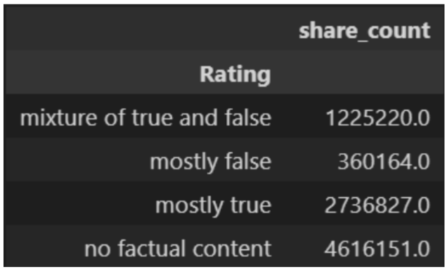
B. Post Removal Below 75% Trust

The second iteration of the beta reputation model calculated the trust values of each user the same way as the first, but this version removed the post if the user’s trust value was below the 75% threshold. According to Figure 5, this iteration showed significantly different results than the two before it. This time, only 21 posts rated “mostly false” were accepted by the model, and only 74 rated “mixture of true and false”. In addition, the shares of “mostly false” posts were reduced by more than half in this iteration, with only 177,070 shares according to Figure 6.



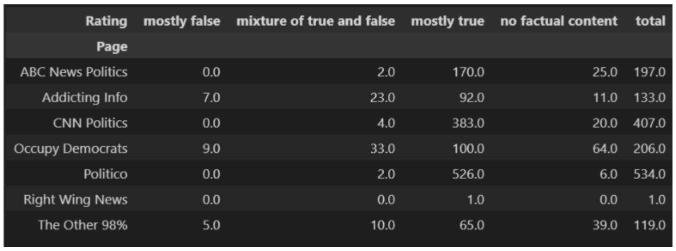
Rating	mostly false	mixture of true and false	mostly true	no factual content	total
Page					
ABC News Politics	0.0	2.0	172.0	26.0	200.0
Addicting Info	8.0	25.0	96.0	11.0	140.0
CNN Politics	0.0	4.0	385.0	20.0	409.0
Eagle Rising	29.0	54.0	121.0	80.0	284.0
Freedom Daily	26.0	26.0	56.0	4.0	112.0
Occupy Democrats	9.0	33.0	102.0	65.0	209.0
Politico	0.0	2.0	528.0	6.0	536.0
Right Wing News	26.0	89.0	142.0	11.0	268.0
The Other 98%	5.0	10.0	67.0	40.0	122.0

Fig. 3. Iteration 1 Approved Posts for Each User



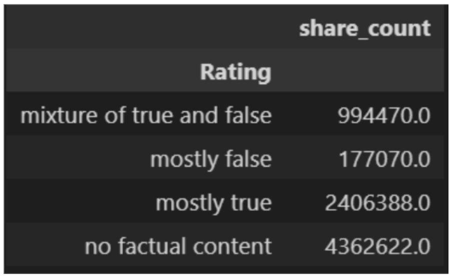
	share_count
Rating	
mixture of true and false	1225220.0
mostly false	360164.0
mostly true	2736827.0
no factual content	4616151.0

Fig. 4. Share Counts After Iteration 1



Rating	mostly false	mixture of true and false	mostly true	no factual content	total
Page					
ABC News Politics	0.0	2.0	170.0	25.0	197.0
Addicting Info	7.0	23.0	92.0	11.0	133.0
CNN Politics	0.0	4.0	383.0	20.0	407.0
Occupy Democrats	9.0	33.0	100.0	64.0	206.0
Politico	0.0	2.0	526.0	6.0	534.0
Right Wing News	0.0	0.0	1.0	0.0	1.0
The Other 98%	5.0	10.0	65.0	39.0	119.0

Fig. 5. Iteration 2 Approved Posts for Each User

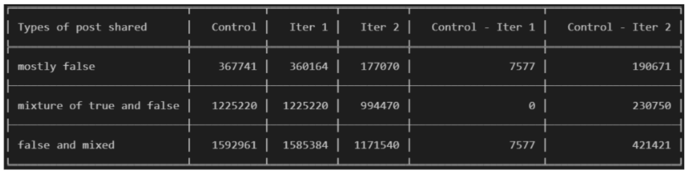


	share_count
Rating	
mixture of true and false	994470.0
mostly false	177070.0
mostly true	2406388.0
no factual content	4362622.0

Fig. 6. Share Counts After Iteration 2

C. Comparing Iterations

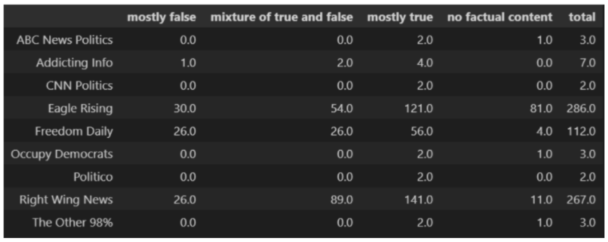
As demonstrated in Figures 1 and 4, there was not a significant difference in the control and first iterations of the beta reputation model. However, the second iteration of the model removed a significant portion of the posts containing misinformation. As shown in Figure 7, the second iteration prevented 51.85% of shares of “mostly false” posts, while the first iteration prevented just 2.06%



```
| Types of post shared        | Control | Iter 1  | Iter 2  | Control - Iter 1 | Control - Iter 2 |
|-----------------------------|---------|---------|---------|------------------|------------------|
| mostly false                |  367741 |  360164 |  177070 |             7577 |           190671 |
| mixture of true and false   | 1225220 | 1225220 |  994470 |                0 |           230750 |
| false and mixed             | 1592961 | 1585384 | 1171540 |             7577 |           421421 |
```

Fig. 7. Comparing Misinformation Share Counts

In addition to the number of shares of misinformation being successfully limited, there were also far more post removals by the second iteration than the first. Whereas the first iteration of the model only removed two posts (one “mostly false” and one “no factual content”), as demonstrated in Figure 8, the second iteration was able to remove 83 “mostly false posts” of the original 104.



	mostly false	mixture of true and false	mostly true	no factual content	total
ABC News Politics	0.0	0.0	2.0	1.0	3.0
Addicting Info	1.0	2.0	4.0	0.0	7.0
CNN Politics	0.0	0.0	2.0	0.0	2.0
Eagle Rising	30.0	54.0	121.0	81.0	286.0
Freedom Daily	26.0	26.0	56.0	4.0	112.0
Occupy Democrats	0.0	0.0	2.0	1.0	3.0
Politico	0.0	0.0	2.0	0.0	2.0
Right Wing News	26.0	89.0	141.0	11.0	267.0
The Other 98%	0.0	0.0	2.0	1.0	3.0

Fig. 8. Posts Removed by Iteration 2

Although the second iteration was able to remove 79.81% of the “mostly false” posts from the network, misinformation was not the only type of post that was removed. According to Figure 8, 332 (19.89%) of “mostly true” posts were also removed.

V. Conclusion

Compared to the control data, removing posts when a user has a trust value below 75% does prevent a significant amount of misinformation from the network, preventing it from being shared further. However, it also removes a significant portion of the truthful information as well. Further research and experimentation are needed in order to find the ideal trust value cutoff for the beta reputation model. With only two posts removed, 25% is much too low, while 75% may be too high as so many truthful posts were removed.

Currently, this model relies on posts already being fact-checked and rated in order to calculate the trust value of the user. Although this means the model is impractical on its own, it could easily be paired with an existing misinformation detection solution. Such a system would be invaluable to social media platforms, as it would be able to effectively classify content as misinformation, quickly remove it as necessary, and analyze a user’s posting habits to determine their trustworthiness.

References

[1] “Russia’s Putin authorised pro-trump ’influence’ campaign, US intelligence says,” BBC News, 17-Mar-2021. [Online]. Available: https://www.bbc.com/news/world-us-canada-56423536. [Accessed: 13-Dec-2021].

[2] Zapan Barua, Sajib Barua, Salma Aktar, Najma Kabir, Mingze Li, ”Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation,” Progress in Disaster Science, Volume 8, 2020, 100119, ISSN 2590-0617,

[3] R. Raju, S. Bhandari, S. A. Mohamud and E. N. Ceesay, ”Transfer Learning Model for Disrupting Misinformation During a COVID-19 Pandemic,” 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 2021, pp. 0245-0250, doi: 10.1109/CCWC51732.2021.9376066.

[4] C. Buntain and J. Golbeck, ”Automatically Identifying Fake News in Popular Twitter Threads,” 2017 IEEE International Conference on Smart Cloud (SmartCloud), 2017, pp. 208-215, doi: 10.1109/SmartCloud.2017.40.

[5] K. Xu, F. Wang, H. Wang and B. Yang, ”Detecting fake news over online social media via domain reputations and content understanding,” in Tsinghua Science and Technology, vol. 25, no. 1, pp. 20-27, Feb. 2020, doi: 10.26599/TST.2018.9010139.

[6] T. Bodnar, C. Tucker, K. Hopkinson and S. G. Bilén, ”Increasing the veracity of event detection on social media networks through user trust modeling,” 2014 IEEE International Conference on Big Data (Big Data), 2014, pp. 636-643, doi: 10.1109/BigData.2014.7004286.

[7] H. Zhang, M. A. Alim, M. T. Thai and H. T. Nguyen, ”Monitor placement to timely detect misinformation in Online Social Networks,” 2015 IEEE International Conference on Communications (ICC), 2015, pp. 1152-1157, doi: 10.1109/ICC.2015.7248478.

[8] A. Maryam and R. Ali, ”Misinformation Source Identification in an Online Social Network,” 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1-5, doi: 10.1109/I2CT45611.2019.9033558.

[9] P. Pourghomi, F. Safieddine, W. Masri and M. Dordevic, ”How to stop spread of misinformation on social media: Facebook plans vs. right-click authenticate approach,” 2017 International Conference on Engineering & MIS (ICEMIS), 2017, pp. 1-8, doi: 10.1109/ICEMIS.2017.8272957.

[10] BuzzFeed (2016) Fact-Checking Facebook Politics Pages [Source code]. https://github.com/BuzzFeedNews/2016-10-facebook-fact-check.