Poster: Detecting Misinformation about Zoom's Security and Privacy Threats

Mohit Singhal, Nihal Kumarswamy, Shreyasi Kinhekar, Shirin Nilizadeh
The University of Texas at Arlington
(mohit.singhal, nihal.kumarswamy, shreyasi.kinhekar)@mavs.uta.edu, shirin.nilizadeh@uta.edu

***Abstract*—Prior works have extensively studied misinformation related to news, politics, and health, however misinformation can also be about technological topics. While less controversial, such misinformation can greatly impact companies' reputations and revenues, and users' online experiences. In this work, we proposed novel approach for detecting misinformation about cybersecurity and privacy threats on social media, focusing on misinformation about *Zoom's security & privacy threats*. Our framework showed great performance with more than 98% accuracy. Running the proposed framework on the posts from Instagram, Facebook, Reddit, and Twitter, we found respectively about 3%, 10%, 4% and 0.4% of posts being misinformation.**

I. Introduction

Prior work has extensively studied misinformation related to news, politics, and health [1]–[3]. Though misinformation can also be about technologies and tools that people use in their everyday life. While less controversial, such misinformation can greatly impact companies' reputations and revenues, and users' online experiences. To detect misinformation about technologies, novel approaches are needed to be employed, as compared to misinformation about news, politics, and health, their purpose and impact are different. One can have roots in people's political and cultural views and beliefs, while the other might take advantage of people's lack of technical background, their beliefs about a technology or company, or their fear about possible security threats.

To the best of our knowledge, no work has systematically studied the spread and characteristics of misinformation about technological topics. In this work, we study misinformation about *Zoom's security & privacy threats* on social media. With the surge in the use of video conferencing tools, such as Zoom [4] during the pandemic, came the concerns about the company handling of security and privacy of its user base. However, not all the discussions were accurate. For example, Tweet [5] claims that Zoom is a "Chinese spying tool." However, the author has not provided any supporting evidence, and the claim is misleading [6].

In this work, we define a post about Zoom's security and privacy threats as misinformation *if the post fails to provide any supporting evidence and it cannot be verified by cross-checking it with reputable sources, or the provided information is fully or partially in contrast with that of trusted sources.*

The presented framework examines the correctness of posts about *Zoom*'s security and privacy threats by (1) obtaining posts from Facebook, Instagram, Reddit, and Twitter, (2) using a ground truth dataset to identify the features that make misinformation posts distinguishable from accurate posts, and (3) using the features to build a classifier that detects misinformation. Our study shows that social media users indeed share misinformation about Zoom's security threats, and that detecting such misinformation is not trivial. We hope that this work increases the awareness of the community and social media platforms about the spread of misinformation about technological topics.

II. Framework for Detecting Misinformation about Zoom's Security and Privacy Threats

In this section, propose a framework for detecting misinformation regarding it on Facebook, Instagram, Twitter, and Reddit. To detect misinformation posts in each social media platform, we developed a binary classifier specific to that platform. Figure 1 shows our proposed framework: (1) Data collection, (2) Groundtruth and taxonomy creation, (3) Feature selection, (4) Training and testing classifiers, and (5) Detecting the misinformation in each platform.

Data Collection & Pre-Processing: In order to collect data from Facebook, Instagram, Reddit, and Twitter, we used the "posts/search" endpoint of the CrowTangle API [7]. We also collected Twitter data using the Observatory on Social Media API [8]. We restricted our data to English and from June 1, 2019, to Nov. 30, 2020. We initially obtained posts that include the keyword *Zoom*. We then employed regex to filter posts and we obtained about 7K, 75K, 9K, and 8K posts for Instagram, Facebook, Reddit, and Twitter respectively.

For training the classifiers, we first manually labeled a subset of the posts (3,300) on each platform to create a groundtruth dataset. Creating this groundtruth dataset, we defined three labels: *(1) Zoom's security and privacy:* if a post satisfies all of the above mentioned criteria, *(2) Misinformation:* if the third criteria is not satisfied, and *(3) Irrelevant:* if it fails to satisfy either first or second criteria. Table I shows the groundtruth dataset that was obtained after the annotation.

Table I: The size of groundtruth datasets per platform.

Platform	Zoom security & privacy	Misinformation	Irrelevant
Instagram	545	15	2,740
Facebook	560	42	2,734
Reddit	1,045	16	2,234
Twitter	1,865	36	1,468

Detecting Misinformation about Zoom's Security and Privacy Threats: We developed a multi-stage classifier to detect misinformation. The first classifier, trains on our groundtruth dataset, and it classifies the remaining dataset into

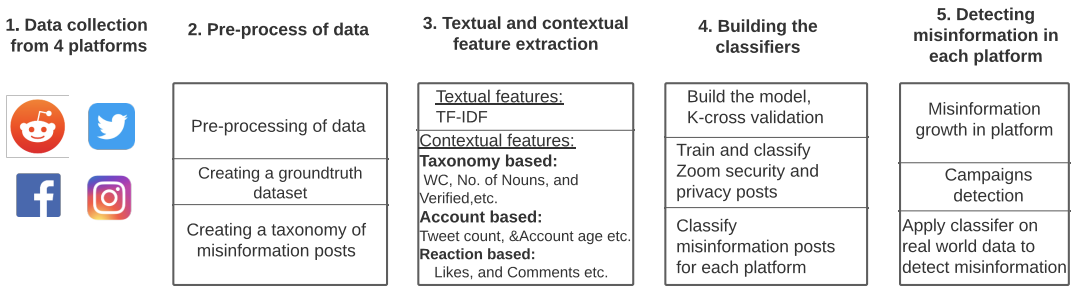


Figure 1: The framework developed for detecting Zoom security & privacy misinformation.

two classes: (1) Zoom security and privacy, and (2) irrelevant to Zoom security. We developed a second binary classifier for each platform, which receives posts related to Zoom's security and privacy and detects if they are misinformation.

Textual Features: For each platform, we extracted bi-grams and uni-grams from all the posts and considered the top 100 of them with the highest values of TF-IDF. **Contextual Features:** We extracted a set of contextual features from the meta-data, which include: (1) *Taxonomy-inspired features:* We created a taxonomy of misinformation about Zoom's security and privacy threats, applying the open coding process [9]. Creating this taxonomy helped us to identify the following features: *word counts*, *noun counts*, *pronouns counts*, *number of all CAPS words*, *misspelled words count*, *verified account*, *followers count*, *has a photo/video* and *has a URL*. (2) *Reaction-inspired features:* Posts can get some reaction. which we used as features: for Instagram, *likes count*, for Facebook, the number of *likes, comments*, and *shares*, for Reddit, the number of *likes*, and *comments*, for Twitter. (3) *Features based on account characteristics:* We also used the following account characteristics can distinguish misinformation posters: *tweets count*, *profile description length*, *account age*, *listed count*, and *has a profile image*.

Classifiers: Since our datasets are unbalanced, we used several oversampling techniques, and found RandomOverSampler provides the best results for Instagram and Reddit classifiers, and SMOTE provides the best results for Facebook and Twitter classifiers. Also, we found that out of the five machine learning algorithms, Random Forest provides the best accuracy across the four platforms. Table II shows all classifiers, using k-cross validation ($k = 3$), have great performance.

Table II: The performance of misinformation detection

Platform	Accuracy	F1 Score	Precision	Recall
Instagram	0.98 (+/- 0.01)	0.98 (+/- 0.02)	0.98 (+/- 0.01)	0.98 (+/- 0.02)
Facebook	0.99 (+/- 0.00)	0.99 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.01)
Reddit	0.99 (+/- 0.01)	0.99 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.01)
Twitter	0.98 (+/- 0.01)	0.98 (+/- 0.02)	0.98 (+/- 0.02)	0.98 (+/- 0.01)

Results: Finally, we employed our model on the posts that are related to security and privacy to detect those that are misinformation. We found that about 3%, 10%, 4% and 0.4% of posts on Instagram, Facebook, Reddit and Twitter are *misinformation*, respectively.

Conclusion: In this work, we proposed a novel approach for detecting misinformation about cybersecurity and privacy threats on social media, focusing about *Zoom's security & privacy threats*. Using a set of textual and contextual features, we built supervised classifiers to identify posts discussing the security and privacy of Zoom, and to detect misinformation in our whole dataset. Our classifiers showed great performance across all four platforms, with more than 98% accuracy, precision and recall. We found about 3%, 10%, 4% and 0.4% of posts on Instagram, Facebook, Reddit, and Twitter, as misinformation, respectively. Our results show that misinformation about cybersecurity and privacy is present on social media platforms and the community needs to further study its impact on end-users and threat intelligence tools.

REFERENCES

[1] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 2931–2937.

[2] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 797–806.

[3] J. S. Love, A. Blumenberg, and Z. Horowitz, "The parallel pandemic: Medical misinformation and covid-19: Primum non nocere," *Journal of general internal medicine*, vol. 35, pp. 2435–2436, 2020.

[4] "The virus changed the way we internet," https://nyti.ms/3Ks3l6D, 2020.

[5] "Stop using Zoom immediately." https://bit.ly/3rlcC7A, 2020.

[6] "PolitiFact," https://bit.ly/33gkCyR, 2020.

[7] C. Team, "Crowdtangle. facebook, menlo park, california, united states," 2020.

[8] C. A. Davis, G. L. Ciampaglia, L. M. Aiello, K. Chung, M. D. Conover, E. Ferrara, A. Flammini, G. C. Fox, X. Gao, B. Gonçalves *et al.*, "Osome: the iuni observatory on social media," *PeerJ Computer Science*, vol. 2, p. e87, 2016.

[9] B. G. Glaser and A. L. Strauss, *Discovery of grounded theory: Strategies for qualitative research*. Routledge, 2017.

Detecting Misinformation about Zoom's Security and Privacy Threats

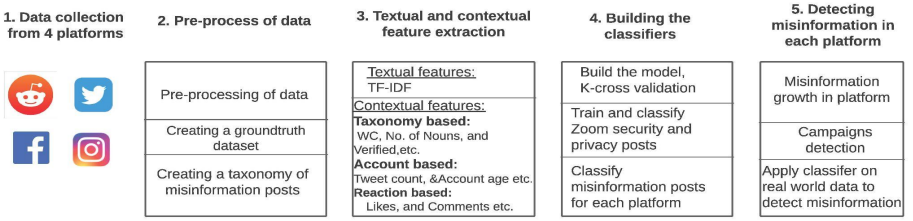
Mohit Singhal, Nihal Kumarswamy, Shreyasi Kinhekar, Shirin Nilizadeh
The University of Texas at Arlington

UNIVERSITY OF TEXAS ARLINGTON | COLLEGE OF ENGINEERING

Motivation

- Prior works have extensively studied misinformation about politics, health, etc.
- To detect misinformation about technologies and tools, novel approaches are needed.
- We propose a novel framework to detect misinformation about Zoom's Security and Privacy threats.

Proposed Framework



Data Collection

- Collected Instagram, Facebook, and Reddit data using Crowdtangle API and Observatory on Social Media API to collect Twitter data from June 1, 2019, to Nov. 30, 2020.
- We then employed regex to filter posts and we obtained about 7K, 75K, 9K, and 8K posts for Instagram, Facebook, Reddit, and Twitter respectively.

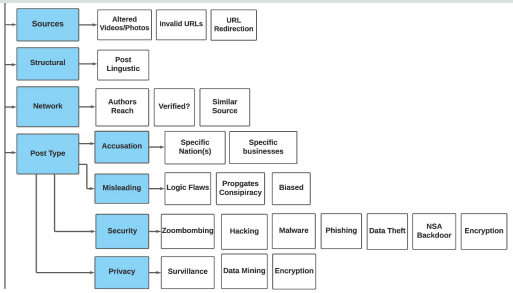
Ground Truth and Taxonomy

- To train our classifier we created a ground truth dataset and we defined misinformation as "*if the post fails to provide any supporting evidence and it cannot be verified by cross-checking it with reputable sources, or the provided information is fully or partially in contrast with that of trusted sources.*"
- Manually labelled 3,300 posts from each platform into three categories.
- We used the misinformation posts to create a taxonomy. The taxonomy yielded features that we used in our classifiers.

Table I: The size of groundtruth datasets per platform.

Platform	Zoom security & privacy	Misinformation	Irrelevant
Instagram	545	15	2,740
Facebook	560	42	2,734
Reddit	1,045	16	2,234
Twitter	1,865	36	1,468

Taxonomy & Misinformation Example



Misinformation Classifier

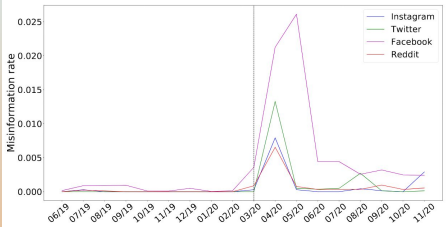
- We developed a multi-stage classifier, the first classifier trains on the groundtruth and classifies the remaining dataset into two classes: Zoom security & privacy posts and irrelevant.
- Our second classifier, detects misinformation posts from the Zoom's security & privacy posts using features from taxonomy, account based and reaction based features.

Table II: The performance of misinformation detection

Platform	Accuracy	F1 Score	Precision	Recall
Instagram	0.98 (+/- 0.01)	0.98 (+/- 0.02)	0.98 (+/- 0.01)	0.98 (+/- 0.02)
Facebook	0.99 (+/- 0.00)	0.99 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.01)
Reddit	0.99 (+/- 0.01)	0.99 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.01)
Twitter	0.98 (+/- 0.01)	0.98 (+/- 0.02)	0.98 (+/- 0.02)	0.98 (+/- 0.01)

- Our classifier showed great performance with more than 98% accuracy across all platforms.

Misinformation Growth



Conclusion

- Proposed a novel framework for detecting misinformation about cybersecurity and privacy threats on social media.
- We built a multi-stage classifier to detect misinformation about Zoom security and privacy threats.
- We found about 3%, 10%, 4% and 0.4% of posts being misinformation on Instagram, Facebook, Reddit, and Twitter respectively.