It's QuizTime: A Study of Online Verification Practices on Twitter

Sukrit Venkatagiri[1], **Jacob Thebault-Spieker**[1], **Sarwat Kazmi**[2], **Efua Akonor**[3], **Kurt Luther**[1]
[1]Virginia Tech,[2]University of Maryland – College Park,[3]Wellesley College
{sukrit, jthebaultspieker, kluther}@vt.edu, skazmi1@umd.edu, eakonor@wellesley.edu
[2, 3] Both authors contributed equally.

Abstract

Misinformation poses a threat to public health, safety, and democracy. Training novices to debunk visual misinformation with image verification techniques has shown promise, yet little is known about how novices do so in the wild, and what methods prove effective. Thus, we studied 225 verification challenges posted by experts on Twitter over one year with the aim of improving novices' skills. We collected, annotated, and analyzed these challenges and over 3,100 replies by 304 unique participants. We find that novices employ multiple tools and approaches, and techniques like collaboration and reverse image search significantly improve performance.

Introduction and Related Work

Today's society is awash with misinformation. Governments and extremist groups spread false propaganda in the form of images and video, which are unwittingly shared by the public on social media. One approach towards combating these misinformation campaigns is *image verification*, a technique employed by expert investigators in domains like journalism and human rights advocacy (Barot, 2014). However, verification is a difficult, time-consuming task. With millions of images shared on social media every day, experts, faced with limited time and attention, are overwhelmed.

Novices, too, attempt to debunk visual misinformation and help investigators (Nhan, Huey, and Broll, 2015; Europol, 2019), but are often hindered by their lack of verification and media literacy skills. Such attempts by crowds of novices can prove erroneous, leading to cases of vigilantism, such as when users on Reddit misidentified a perpetrator of the Boston Marathon Bombing (Nhan, Huey, and Broll, 2015).

Researchers have studied ways to better understand expert verification practices (Brandtzaeg et al., 2016). Other work seeks to provide software tools to help novice crowds support expert investigators on verification tasks, such as image geolocation, the goal of which is to find the exact location on earth where a photo or video was taken (Kohler, Purviance, and Luther, 2017; Venkatagiri et al., 2019). Prior work (e.g., Caulfield, 2017) has studied ways to improve novices' media literacy and verification skills. However, they were primarily controlled lab experiments, and focused on outcomes and not process (Venkatagiri and Zhang, 2018). Little is known about how novices learn verification skills in the wild, what tools and approaches they use, how successful they are in verifying media, and what factors affect success.

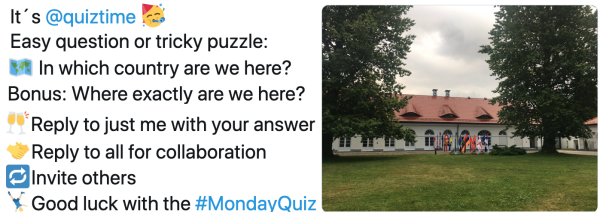

Figure 1: A representative example of a verification quiz.

Here we attempt to answer these questions by studying real-world, collaborative verification challenges (quizzes) posted on Twitter over the span of one year. We scraped all quizzes and associated conversation threads, and annotated each quiz for descriptive data and participants' performance. We also calculated engagement metrics, such as the total number of replies and total number of unique users.

Our preliminary analysis found that (1) novices employ a variety of tools and creative techniques when performing image verification; (2) a majority of participants were successful in doing so; and (3) whether they collaborated on a solution, or used certain tools, meant they were statistically significantly more successful. We conclude with a discussion of our results and avenues for future work.

Methods

Background and Data Collection. These quizzes aim to help novices practice and improve their skills (Faure, 2019). An expert posts a quiz from their personal account with the specific hashtag once per weekday (e.g., #MondayQuiz), and are retweeted and aggregated by the Verification Quiz Bot (VQB) Twitter account[1]. After an expert posts a quiz, they will reply to the tweet, which serves as a separate thread that acts as a "spoiler barrier" for participants to reply with their answers, and for the expert to verify. Outside of this thread, participants can collaborate to solve the quiz, or ask for further details. We systematically scraped all quizzes posted during 2018 that the VQB had retweeted, and those

[1]https://twitter.com/quiztime

with the associated hashtag for each day's quiz. After sorting and cleaning, there were 225 daily quizzes for 2018.

Codebook Development. Two of the authors inspected a set of 10 quizzes and replies to develop a codebook, which was iteratively refined. After the final iteration, the codebook had three categories for the daily quizzes: (1) the type of question posed in the quiz (e.g., who, what, when, where, refute/verify, number of hints); (2) the type of media associated with the question (photo, video, audio); and (3) the content of the media itself (architecture, cityscapes, landscapes, aerial/satellite imagery, objects/animals, people, other). We also determined the ground truth answer for each quiz. The codebook for replies to each quiz had three categories: (1) response type (an answer or details); (2) process type (tools used, use of deductive reasoning, having been there); and (3) if there was an answer, whether it was completely correct (C), partly correct (PC), or incorrect (IC).

Data Coding and Analysis. After finalizing the codebook, two authors coded the same subset of 50 quizzes and replies. They then met to resolve disagreements in interpretation, and then the third author continued to code the remaining 175 quizzes and associated replies. After the coding process, we calculated metrics based on tweet metadata, such as the time duration of each quiz and the number of replies per user. Initial observations during the coding process did not provide any preliminary evidence that factors like the use of Google Maps or using tools beyond reverse image search (RIS) showed any differences in performance. Thus, we did not conduct hypothesis tests for these factors. Instead, we found that people tended to perform better when they used RIS, collaborated, used deductive reasoning, and hints were provided. To measure performance, we assigned replies a score of 2 if the answer was completely correct, 1 if partly correct, and 0 if incorrect. We conducted follow-up statistical analyses in R. We used Mann-Whitney U tests as a non-parametric alternative to the two sample t-test to compare performance between groups (see *Improving Performance*).

Preliminary Findings and Discussion

Quiz Question. The most common question was to identify the location depicted in the media for each quiz (202/225, 90%), followed by object identification (37%). The least common question involved identifying people in or associated with the media (14%). There was often more than one question asked per quiz, with 533 total questions over 225 quizzes, e.g., "*The historical heritage of the Balkans is impressive, like this abandoned temple. (1) Where is it? (2) When was it built? (3) Am I telling the truth?*"

Quiz Media Content. There was a total of 207 photos, 11 videos, and 7 audio clips. The most commonly depicted visual elements were architecture, buildings, and structures (168/225, 75%), followed by objects (42%), and landscapes (42%). The least common was satellite imagery (4%).

Participation. There was a total of 3,100 replies within the 225 quizzes, including those by the quizmasters (mean = 13.8, s.d. = 5.8). There were 304 unique participants, who replied directly to the quiz tweet between 1 and 67 times (mean = 3, s.d. = 7). 294 unique participants made at least one attempt to answer a quiz. The 9 quizmasters accounted for 41% of replies, helping other participants as well as being participants in others' quizzes.

Performance. Of the 3,100 total replies, 891 (29%) contained an answer. Of these answers, 60% were completely correct, 27% were partly correct, and 13% were incorrect. The low proportion of incorrect answers reflects verification as a "Eureka"-style problem where an answers' correctness is immediately obvious. Thus, participants with wrong answers likely realized as such and did not reply.We found no evidence that time taken ($\log_{10}$-scaled) and performance were correlated (r=0.06, p-value=0.41). We also found no trend between quizmaster participation and overall performance ratios for each quiz.

Approaches. 671 of 3100 replies provided details on tools and methods that people used to arrive at an answer. Google Maps was the most commonly used (33%), followed by other tools and methods (30%), inspecting photo metadata (24%), using search engines (22%), and deductive reasoning (22%). Surprisingly, reverse image search (RIS) tools were mentioned in only 13%. 6% of the replies with details involved users collaborating to arrive at an answer. Only 3% of replies involved an answer where the participant indicated that they had previously seen it in-person or online. 51% of replies utilized two or more approaches, e.g., "*The licence plate pointed to a location in North Rhine-Westphalia. I did a Google search for Cologne + bridge + grafitti. It brought up a different Joiny picture... I then scrolled through some websites about Joiny and found the right one.*" Please see Supplementary Materials for further details and examples.

Improving Performance. While RIS tools—e.g., Google, Yandex, and TinEye—were mentioned 13% of the time, they were always used in conjunction with another technique. A Mann–Whitney U test found that the use of RIS tools had a significant effect on performance (mean = 1.83), compared to other methods (mean = 1.47) ($W = 19351$, $p < 0.001$). User collaboration also had a significant effect on performance (mean = 1.84) compared to those who did not collaborate (mean = 1.45) ($W = 20671$, $p < 0.001$). We found no significant difference in performance based on whether users used deductive reasoning (mean = 1.44) or not (mean = 1.47) ($W = 50978$, $p = 0.52$), or whether a quiz had a hint (mean = 1.44) or not (mean = 1.48) ($W = 93266$, $p = 0.48$).

Discussion and Future Work

Our findings suggest that media verification requires one to geolocate media 90% of the time. This is in line with prior work that has shown that geolocation is crucial to verification (Barot, 2014). Our findings also suggest that novices employ a number of verification tools and techniques, often in a pipeline. We also find that certain tools and techniques were underused—such as RIS and collaborating with others—but proved fruitful for those who did. Future work should explore studying a larger number of quizzes, running further statistical tests to see what factors affect performance, such as a temporal analysis; and exploring multiple regression models with input parameters such as number of tools used, and whether there was collaboration. Future work should also explore semi-automated pipelines to support novice verification practice, and not just outcomes.

References

Barot, T. 2014. Verifying images. In *Verification Handbook: A Definitive Guide to Verifying Digital Content for Emergency Coverage*.

Brandtzaeg, P. B.; Lüders, M.; Spangenberg, J.; Rath-Wiggins, L.; and Følstad, A. 2016. Emerging journalistic verification practices concerning social media. *Journalism Practice* 10(3):323–342.

Caulfield, M. 2017. *Web literacy for student fact-checkers*. Press Books.

Europol. 2019. Stop child abuse – trace an object.

Faure, G. 2019. The Daily Quiz That Teaches Journalists How to Geolocate Images. *Global Investigative Journalism Network*.

Kohler, R.; Purviance, J.; and Luther, K. 2017. Supporting image geolocation with diagramming and crowdsourcing. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

Nhan, J.; Huey, L.; and Broll, R. 2015. Digilantism: An Analysis of Crowdsourcing and the Boston Marathon Bombings. *The British Journal of Criminology* 57(2):341–361.

Venkatagiri, S., and Zhang, A. X. 2018. Response to "heuristics for the online curator". 9–10. Research Triangle Park, NC: RTI Press.

Venkatagiri, S.; Thebault-Spieker, J.; Kohler, R.; Purviance, J.; Mansur, R. S.; and Luther, K. 2019. GroundTruth: Augmenting expert image geolocation with crowdsourcing and sared representations. *Proceedings of the ACM on Human-Computer Interaction* (CSCW 2019).