Apr 20th, 12:00 AM - 12:00 AM

Extracting Information from Twitter Screenshots

Tarannum Zaki
Old Dominion University

Michael L. Nelson
Old Dominion University

Michele C. Weigle
Old Dominion University

EXTRACTING INFORMATION FROM TWITTER SCREENSHOTS

Tarannum Zaki[1], Michael L. Nelson[1], and Michele C. Weigle[1]
[1]Old Dominion University, Norfolk, VA, USA
E-mail (tzaki001@odu.edu)

ABSTRACT
Screenshots are prevalent on social media as a common approach for information sharing. Users rarely verify before sharing screenshots whether they are fake or real. Information sharing through fake screenshots can be highly responsible for misinformation and disinformation spread on social media. There are services of the live web and web archives that could be used to validate the content of a screenshot. We are going to develop a tool that would automatically provide a probability whether a screenshot is fake by using the services of the live web and web archives.

Keywords: Twitter, misinformation, disinformation, screenshot, web archives.

INTRODUCTION
A screenshot is a way to share content on social media that allows cross-platform user engagement. For example, @RBReich shared his own tweet as a screenshot on Facebook to increase cross-platform engagement (Reich, 2022). Moreover, screenshots are also used as evidence where there are chances of controversial posts getting deleted. For example, a tweet posted by @DanielDefense about the Uvalde shooting incident was deleted but an archived version exists (Defense, 2022). Fake tweets are relatively easy to create using online tools such as Tweetgen[1]. It is difficult to edit or delete a screenshot of a fake tweet once it has been shared across social media. So, it becomes quite challenging to detect whether the content of the screenshot is real or fake.

There are no tools currently available to evaluate the authenticity of screenshots shared on social media. But, there are methods which could be utilized to verify whether the content of a screenshot had been really posted by the author. The simplest way of searching on the live web is by searching for the text in a search engine, such as Google. Furthermore, there exist fact-checking websites, like Snopes[2] and FactCheck.org[3], where users can search whether content is fabricated. For example, a fabricated tweet by Rep. Marjorie Taylor Greene regarding 4th of July was fact-checked by FactCheck.org (Spencer, 2022). Another useful method in this regard is searching web archives such as the Wayback Machine[4] for deleted posts or accounts. The previously mentioned example of @DanielDefense is an example of finding deleted tweets on web archives.

So, there are ways to validate whether the content of a screenshot is real or fake by utilizing the live web and web archives. We are developing an automated tool using such services to estimate the likelihood that the content of a shared screenshot on social media had been really posted by the alleged author. The contribution of our research would definitely aid in alleviating misinformation and disinformation spread by detecting whether the content of a screenshot is fake or real.

MATERIALS AND METHODS
First, we created a data set of screenshots posted on Twitter. The data set currently consists of 200 screenshot images with both real and fake examples. A detailed description of the data set is provided in our blog post (Zaki, 2022).

The next task is to backtrack the original link of the content of the screenshot. This would be done by searching the live web using methods like Google search and fact-checking websites. This requires the tweet text to validate the screenshot. Another way is to search web archives such as the

[1] https://www.tweetgen.com/
[2] https://www.snopes.com/
[3] https://www.factcheck.org/
[4] https://archive.org/web/

Wayback Machine. This requires the timestamp and Twitter handle from the screenshot. The CDX API[5] of the Wayback Machine helps to retrieve a tweet's URL using this information.

Finally, if content had really been posted by the alleged author, there is the possibility of finding it on the live web and web archives. This would help to determine whether the screenshot is valid.

RESULTS AND DISCUSSION

The input for the intended tool are screenshot images. Initially, optical character recognition (OCR)[6] is applied to extract the information available in the screenshot. Our methods require the extraction of tweet text, timestamp, and Twitter handle. So far, we have performed an experiment to evaluate two methods for extracting the timestamp using 125 single tweet images from our data set. Method 1 uses the Python module datefinder[7] for extracting timestamp, whereas Method 2 uses an additional logic on the date format along with it. Table 1 shows that Method 2 performs better than Method 1 in terms of accuracy, precision, recall, and F1 score.

Table 1. Performance evaluation of methods for extracting timestamp.

Methods	Accuracy	Precision	Recall	F1 Score
Method 1	41%	60%	39%	47%
Method 2	80%	74%	97%	89%

CONCLUSION

Screenshots are the easiest way to share content on social media. As there does not exist any specific tool to establish veracity of content that is shared as a screenshot, it is a critical task to detect a fabricated post. It is important to verify whether a post is fake or real, because fake posts can be responsible for misinformation and disinformation spread on social media. The collected data set and the tool we are developing for this research would greatly contribute to the research areas of misinformation and disinformation spread on social media.

ACKNOWLEDGMENT

This work is supported by the GROW M&S project (Grant # 300747-010), funded by the US Department of Education.

REFERENCES

Reich, R. [Class in Session]. (2022, Aug 18). *What do you think happens if Trump runs in 2024* [Facebook page]. Facebook. https://www.facebook.com/watchclassinsession/posts/pfbid0344Hu2bxJtAiiL5VHfM2YQyPTU9jTm3tfdJMj4TZMDunomMarXMQfTxPGvsVwfBmwl

Internet Archive. [@DanielDefense]. (2022, May 16). *Train up a child in a way he should go* [Tweet]. Twitter. https://web.archive.org/web/20220525125749/https://twitter.com/DanielDefense/status/1526237750277681154

Spencer, S.H. (2022, July 5). *Fabricated Fourth of July Tweet Was Not from Rep. Marjorie Taylor Greene*. https://www.factcheck.org/2022/07/fabricated-fourth-of-july-tweet-was-not-from-rep-marjorie-taylor-greene/

Zaki, T. (2022, Dec 12). *Disinformation Spread on Social Media through Screenshot Sharing: Dataset Description*. Web Science and Digital Libraries Research Group. https://ws-dl.blogspot.com/2022/12/2022-12-12-disinformation-spread-on.html

[5] https://archive.org/help/wayback_api.php

[6] https://en.wikipedia.org/wiki/Optical_character_Recognition

[7] https://pypi.org/project/datefinder/