Is it real? A Social Media Post Evaluation

Mukul Prabhu
40257131

Abstract

Social media platforms are the news reporting platforms of this age. It started off with newspapers, radio, and television but the spread of social media has provided an accidental feature. Instant news feeds at our fingertips. But unlike radios, papers and television, social media has very little regulation in comparison. If we take news channels, they have a certain reputation to uphold and must have a fact-checking process to avoid giving false information to the people. But that's not the case in social media, here any person or even a bot can create an account and post altered images. There are guidelines and policies present to prevent such situations but by the time companies take action the damage has already been done. Photoshop has always been a prime tool used to alter images but with the rise of generational AI, the lines between reality and fantasy have blurred even more. There are many ways to spread false information such as altered images, videos with a different context, voice impersonation and many more. To streamline the research, this project focuses on the prevention of false information spread through AI-generated videos on Instagram. Current methods used to fight against misinformation will be analysed and see how they can be used against AI-generated videos. An evaluation framework table will also be used for the same. The paper will also dwell on a possible solution of its own as well.

Introduction

In simple terms, fake news can be defined as a news short that deliberately misinforms the public of current events. False information has always been present in social media but the general public started being more serious about such misinformation with the beginning of COVID-19. During the peak infestation time, social media posts were stating that COVID was spread through 5G antennas and some people believed it too. The WHO had to deal with not only the virus but also several false information spreading around social media.

This was a simple example, the situation gets complicated as we start to look into deformed images. Photo modification applications are easily accessible and can be used to change as per the user's wish. There have been some recent events of such use as well. On Mother's Day, the princess of Wales Kate Middleton posted a picture along with her kids which she later confirmed had been touched up by herself using Photoshop. An important note here, a lot of people on social media started pointing out inconsistencies with the image meaning a certain level of skill is required to pull it off and as she had done the editing herself, there were no ill intentions behind it. In another case, there was a string of AI-generated explicit images of Taylor Swift posted on X (Twitter). This clearly was a job done by a third party. And so, the situation has now changed. There's no need to have a special skill to modify images anymore, a user can type in a prompt and AI will generate the image for us. Nowadays with AI, it has become even more difficult to figure out if the image is real or fake. With the announcement of OpenAI's DALL-E 3, it can be close to impossible to detect if it is fake with the naked eye. Now OpenAi has also introduced a video generation tool called Sora.

One obvious harm such images produce is on the person's reputation. To protect said reputation social media platforms can use their own technology to detect morphed images, but what about realistic-looking videos? Is the current system capable of handling this new technology? Can an independent user figure out if the video is generated artificially? Even though this new technology will quickly be available to everyone, the means to safeguard against it is not known.

Existing Research

Paper 1: Testing the Effectiveness of Correction Placement and Type on Instagram
Based on climate change posts

This paper tries an experimental way of correcting misinformation using fact-focused and logic-focused corrections also based on the placement i.e. before vs. after the misinformation message. Instagram has its own ways to fight misinformation but this paper explores ways that users can do to confront misinformation. For pre-bunking (facts before misinformation), the framework first shows a warning regarding the prevalence of misinformation or a technique used to mislead audiences and then exposes users to misinformation. For debunking, it shows factual corrective information after exposure to misinformation.

First, as a general rule, debunking strategies have largely focused on facts whereas prebunking efforts may focus either on facts or on logic. The paper conducted an experiment where volunteers were asked to go through a controlled feed on Instagram posts with one of them being false and deployed various combinations of logic-based and fact-based corrections along with prebunking and debunking corrective measures. Overall, it was found that logic-focused corrections outperform fact-focused corrections. While the logic-focused and fact-focused corrections are equally effective when seen after the misinformation post, only the logic-focused corrections reduce misperceptions when they appear before the misinformation, offering more flexibility in their application.

First, they function as a proactive strategy; one does not need to wait until one sees misinformation being spread to correct it. It demonstrates that corrections need not directly reply to misinformation, so long as they appear immediately afterwards (or before, when using a logic-focused approach). It is impossible to know exactly when or where people have seen misinformation. The best approach was providing a fact-based prebunking, but the Instagram algorithm wouldn't always have this fact-based post placed before the actual misinformation post, a better way would be to tie this information with the false post itself. Applying this to a video, a logic-focused post can make the user think about the possibility of the video being real. Making the user doubt the content and not blindly believe it is a good approach as long as that logic-based post is seen before the video.

Paper 2: Fact-checking Literacy of Covid-19 Infodemic on Social Media in Indonesia

This paper explores fact-checking focused on COVID-19. The paper analyses posts by the account "@jabersaberhoaks" a government of Indonesia initiative. The posts are

fact-checked by checking official sources such as media, authorised agencies, and expert sources nationally and internationally. The police tried to handle the spread of fake news by announcing the chief of police's declaration, taking help from their COVID-19 response service, cyber patrol and socialising this issue on social media platforms. The paper deals with finding out how the fact-checking process works, indicating that if people know about the underlying process it's easier for them to trust it.

Content Analysis was used to determine if the posts were fake. The process referred from the Handbook of Journalism and Training Education consists of finding fact-checkable claims on the post, finding those facts and correcting the record. This solution is based on having a trusted channel that will post verified content. But this doesn't stop other accounts from spreading false information nor does it help in indicating to people that the content itself could be false. Now the way this channel verifies information is by going through making comparisons with reports from trusted media sources at national and international levels through news agencies and verified media sources, including legitimate journals. They go through independent fact-checking websites and tracking statements from experts as well. The positive aspect of this is that the government has taken the initiative to fact check information and the people don't have to got through countless sources to verify the same. The downside is that it takes a lot of time to process this. This issue was solved using advanced machine learning models for natural language processing.

Say in a particular situation there are posts put up on Instagram giving regular updates but as the situation progresses the previous posts could end up contradicting to what is the current scenario, and these old posts could randomly be shown to a user. Unless the account tries to go through its previous posts and delete outdated information, it will do more harm than good. Also there can be a case where a dispute between 2 countries generates alot of posts and both country's official accounts post contradicting information on the same topic. For the local people who are involved and present in the situation could tell who is lying but it's difficult for everyone else in the world to trust these posts in spite of being from official government accounts.
Note: At the moment that account is not available on Instagram but it is present on Facebook.

Paper 3: "See the Image in Different Contexts": Using Reverse Image Search to Support the Identification of Fake News in Instagram-like Social Media

This method tries to find fake news by utilising a web-based companion called virtual learning companion (VLC) that can use reverse image search to show the same image being used in different contexts. It uses natural language processing to provide keywords from other sources to the user to make informed choices. It also asks the user some questions based on the scanned posts making the user part of the verifying process. One example could be a post of an accident reported to have happened early in the day but turns out the image was reused from a previous incident that happened in some other country altogether.

The paper uses PixelFed to simulate an Instagram-like environment and the virtual learning companion is set up through an extension. In practice, the user would need to right-click on the post and send its URL through the plugin. The extension then uses its reverse image

search to find possible matches. The chatbot in turn asks the user to rate the post as fact or fake and then asks for follow-up questions based on the user's choice. The chatbot then gives the user possible matches from its reverse image search along with the source URLs.

Unlike the previous paper, here the user is asked to determine if what is shown is true or false by providing related sources scraped from the web. By showing sources having the same image but in a different context helps users to make an informed decision, this also bypasses the possibility of a "trusted" entity not being honest when informing people of a fake post. Basic natural language processing is also used to extract any text from the image that will help in finding related sources. This methodology seems to be the most effective and as the user is provided with many sources, it is all the more believable for them. It seems that this is similar to just exiting Instagram and doing a Google search but this automated process saves time and also selects relevant and trusted sources from the web instead of the user stumbling across a site that ends up promoting the fake post instead.

Paper 4: Ibero-American fact-checkers on Instagram: analysis of the posts with the most successful interaction

This paper focuses on the fact-checkers themselves. It details what type of content they post, their consistency, and credibility, and also the workings of Instagram's backend. One of the main reasons Instagram has become a hotspot for news activity is the speed at which information reaches people or rather gets "viral". This depends not only on the follower base but also on how often the users post on the account and how many people engage with the content. This is important for fact-checkers as they want their verified or debunked content to be available to the people in their feeds. As mentioned in Paper 1, it is better to have people see verified content before the fake one, this can be achieved if the fact-checkers account is engaging and famous.

Another aspect to look upon is the "algorithm". For the general users, it is this unexplainable program that delivers posts from different content creators. Even for people who are part of the tech industry, this algorithm is still a mystery as Instagram hardly ever discloses how it works. It's like playing a game without knowing the rules. But one characteristic is known by all, if a content is viral (like a song) and a user uses it in their posts, the algorithm will deliver it to the people. This is how news spreads so quickly, users will use a famous song or a humorous meme format to get as many views as possible cause that content is viral at the moment and is receiving the most user engagement.

Fact checker have their own limitations. It is not possible for them to verify every story or fake news that gets published. The paper surveyed a group of fact-checking accounts and found that not all posts are related to fact-checking. Some are information-based and others are self-promotion. Also, every organisation has a bias towards a particular topic where they would focus more on. There are accounts that focus more on politics while some on health. The reason is to increase user engagement and work on topics that trending. It should be noted that of all the content posted by the surveyed accounts, just 13.3% of them were videos meaning, fact-checkers post more image-based or information-based content. The reasons are not discussed but it could be that it's more time-consuming or complicated to verify video-based content. There is also the engagement factor. Through their analysis, it

was found that very few videos generated the level of engagement i.e. likes compared to their image and information counterparts.

Present Day Process

Going through Instagram's policies and community guidelines, we will now see how these affect misinformation keeping in mind the project's focus on AI-generated videos. Instagram has provided some tips for identifying false information such as verifying the video's source by searching "elsewhere" - they did not provide any reliable source to cross-check. Checking for the date of the post, being sceptical of bold claims made in the post, and checking for unusual formatting. These tips largely apply to images and aren't very effective for AI-generated videos. Instagram states that to identify false information it takes the help of its users by asking for feedback and working with third-party fact-checkers. The process is initiated by the user or the fact checker is free to identify and verify content on its own. There are no clear selection criteria for verifying a particular domain of misinformation. Also many people would've seen or engaged with the post before it actually receives user feedback.

When a post is rated false, partly false, or missing context by third-party fact-checkers Instagram reduces its visibility and makes that post harder to find and is marked by labels but it's not deleted until it goes against the community guidelines. The community guideline has rules for intellectual property rights, nudity, spam, impersonation, harassment, organised crime, pornography, racism and much more. Instagram can take down the offender's post and even terminate the account if found guilty but there is no mention of any legal action against them. If we consider this statement by Instagram on their content moderation page, "If content doesn't go against Community Guidelines but may be inappropriate, disrespectful, or offensive, we may limit it from Explore, rather than removing it from Instagram", it leaves a lot of room for discussion as what can be considered as offensive, that said an AI-generated video can still cause enough damage.

Instagram follows a model where it prioritises quantity over quality. Action or investigation of a post is done only after users send feedback, content is deleted only in extreme cases and the rest of the posts are limited. In terms of accountability, the user's account can be terminated but that's the worst-case scenario. This type of model welcomes trouble first and then resources are spent trying to control the damage done. One possible solution could be content screening before the platform publishes it. And the platform could display a badge saying its verified information. This approach gives more importance to security and can be a downside in terms of usability and deployability.

Evaluation Framework

The framework will evaluate the current methodologies present to tackle misinformation and try to see how they can handle the emerging threat of AI-generated videos. The methodologies will be used from the papers described in existing research, Instagram's current process and also the model suggested in the previous section. It is assumed here that these systems have a way of distinguishing between an ai generated video and a real one.

Category	U1	U2	U3	S1	S2	S3	D1	D2	D3
Use of correction post	◉	◉	◉	◖	◖	◉	◖		◉
3rd party fact-checkers	◖	◖		◖	◖	◉	◖	◉	◉
Web extension or plugin	◖		◖	◉	◉	◖	◖	◖	◖
Hybrid model by Instagram	◖		◖	◉	◉	◉		◖	◖
Pre-screening posts	◉	◖	◖	◉	◉	◉			

Usability Criteria

- U1: Ease of Finding Facts
    - Full Dot: Available immediately
    - Half Dot: Need to make a search
    - No Dot: Not easy
- U2: Easy to Use System
    - Full Dot: Very Easy
    - Half Dot: Takes a couple of steps to get the result
    - No Dot: This can be complicated
- U3: Using System to Provide Feedback
    - Full Dot: Can be done quickly
    - Half Dot: Takes a couple of steps
    - No Dot: Takes too long (User might not do it again)

Security Criteria

- S1: Resilient to obtaining wrong information
    - Full Dot: Won't provide wrong information to the user
    - Half Dot: The user will need to judge if the information is true
    - No Dot: Possible of getting wrong information
- S2: Trustfulness of the system
    - Full Dot: The system is trustful
    - Half Dot: The user will need to judge if the information is true
    - No Dot: The system cannot be trusted
- S3: Resilient to exposing user data
    - Full Dot: The user's data is secure
    - Half Dot: Chance of user exposing data
    - No Dot: User's data is exposed

Deployability Criteria

- D1: Quick processing
    - Full Dot: The platform can provide results immediately
    - Half Dot: The platform will take a few minutes before showing results
    - No Dot: It takes a long time to see results
- D2: Easy Implementation
    - Full Dot: It's easy to implement

    - Half Dot: Not difficult but requires some expertise
    - No Dot: It’s difficult to implement
- D3: Cheap
    - Full Dot: Doesn't cost much to implement
    - Half Dot: Not expensive but requires some amount
    - No Dot: Expensive to implement

Conclusion

This paper demonstrates the present methodologies available to fight misinformation on the social media platform Instagram, and if these methods are still viable when it comes to AI-generated videos. The reason for focusing on videos is that there are tools available to detect AI-generated images, but not for videos. Those tools fail if used on video-based content will fail. We are then dependent on companies to have a type of indication on the video like a watermark to indicate that it’s AI-generated. But adversaries won’t be doing it and thus there’s a need to figure out if the video is trying to impersonate a real person or if the person’s dialogue in the video is faked (example: presidential campaigns).

The methodologies used in the evaluation framework have the right base. Most of the processes can be used for video-based content but with the assumption that these systems have access to the appropriate detection tool. Though the proposed system seems to be ideal in terms of security, it is difficult to imagine Instagram would consider opting for it as pre-screening content can be time-consuming and could generate a lot of false positives. A paper that is scheduled to be presented in June 2024 called “Beyond Deepfake Images: Detecting AI-Generated Videos” (Reference Paper 5) figured out a way to distinguish real and AI-generated videos. The authors have trained a model to detect distinct forensic traces left by these video generators and can detect other videos by the same generator as well. Though if a video of a new generator is used, the results aren't satisfactory but once trained on a few videos the detection percentage goes up. If a similar technology is provided to developers, it’s possible to integrate it into the current processes.

However, at the moment, users need to brace themselves as this type of technology is not available at the moment and AI-generated videos are already popping up on social media. One way to help fight this would be to have some legal aid to keep this situation under control. The United Kingdom recently passed a law that made it a criminal offence to create explicit videos or images. Users should also be vigilant and not believe what is shown on social media. It’s always a good practice to fact-check if a post makes any bold claims and not share it. Looking out for misinformation labels and trusting official sources should be always done.

References

News Article: Kate Middleton’s edited Mother’s Day photo, explained by an expert

News Article: Trolls have flooded X with graphic Taylor Swift AI fakes

News Article: Instagram's policy changes

News Article: UK passes law against explicit images generated through deepfake

Paper 1: Testing the Effectiveness of Correction Placement and Type on Instagram

Paper 2: Fact-checking Literacy of Covid-19 Infodemic on Social Media in Indonesia

Paper 3: See the Image in Different Contexts”: Using Reverse Image Search to Support the Identification of Fake News in Instagram-Like Social Media

Paper 4: Ibero-American fact-checkers on Instagram: Analysis of the posts with the most successful interaction

Paper 5: Beyond Deepfake Images: Detecting AI-Generated Videos

Evaluation Framework Reference: The quest to replace passwords - a framework for comparative evaluation of Web authentication schemes

Instagram: Tips for identifying false information

Instagram: Reducing spread of false information

Instagram: Terms of Use

Instagram: AI in content moderation

Instagram: Community Guidlines