Quantifying the Emergence of Deepfakes

1st Kaushal Bhat
Computer Science
University of Arizona
Tucson, Arizona USA
kaushalbhat@arizona.edu

2nd Anthony Bisgood
Computer Science
University of Arizona
Tucson, Arizona USA
anthonybisgood@arizona.edu

3rd Carlos Field-Sierra
Computer Science
University of Arizona
Tucson, Arizona USA
fieldsierra@arizona.edu

4th Joseph Clancy
Computer Science
University of Arizona
Tucson, Arizona USA
clancy1@arizona.edu

5th Chase Klapperich
Computer Science
University of Arizona
Tucson, Arizona USA
chaseklapperich@arizona.edu

***Abstract*—Deepfakes often have malicious intent and are a growing problem. We developed three methods of quantification of the growth of deepfakes on social media to reveal information about the deepfake problem that may be useful for developing detection software.**

I. Introduction

Following the recent explosion of advancements in artificial intelligence , deep learning models, including those in the GPT family, have made giant strides in their efforts to imitate and generate synthetic and deepfaked text [2]. While there have been some advantageous purposes for generating texts that can be passable as human-written, from entertainment and joke creation to text summarization [11], just as any other tool can be used righteously or for wicked purposes, deepfake text generation is another example of a double-edged sword.

While in the past, a single human might have been capable of producing and disseminating misinformation online, unchecked access to the generation of deepfake texts could entail the production of misleading facts and figures at an unprecedented speed and scale. A bad actor with the wrong intentions could potentially spread false information, propaganda, or manipulative content at a much higher rate and to a much larger audience than what was ever possible before [11]. Furthermore, since the content itself would not be produced by a human, there would be no clear figure to hold accountable for the repercussions of the content. While human-generated content might have the potential to also be harmful, the characteristics of artificially generated synthetic text make for a problem that is uniquely challenging to both identify and mitigate.

Deepfakes are becoming an increasingly common presence on social media sites. They are a security concern because they can be very realistic and have an influence on a large number of people with misleading or offensive content. Because of this, deepfake detection software has become an important focus to researchers. Development of deepfake detection requires information about the deepfake problem on social media. Twitter is one of the most popular social media websites and is known to have a problem with accounts posting deepfake content. This is why our project tries to answer the question: how best to quantify the emergence of deepfakes on Twitter?

To answer our research question, we developed three methods of quantification of textual deepfake content on Twitter that may be used to aid development of deepfake detection software: engagement over time, network graph, and region. The first quantifies the growth of deepfake accounts by their engagement over time. We sampled various metrics of engagement over time at a certain interval. The second method was planned in order to reveal information about the relationships between deepfake accounts internally and externally. That is, are they mostly interacting with each other or real users? Our third method collects data about the deepfake accounts' regions to see if any regions have a higher amount of deepfake accounts relative to ot. All this data could potentially reveal information about the growing deepfake problem that could be useful when developing methods or heuristics to detect deepfakes.

II. Design and Implementation of Systems

We propose 3 different ways to best understand and quantify the emergence of deepfakes on social media networks. One method measures the emergence of deepfakes by using density graphs between deepfake accounts. The other measures user engagement (likes, comments, reposts) from accounts that post deepfakes and their growth. The last method measures deepfake accounts growth to their respective geographic region. Each of these methods was done using web scraping. To detect if content was deepfaked, we used an API tool called "AI content detector" that allowed us to input text and returned a detection score of how "human generated" the text was. If the text was less than 50% human generated, we considered the text deepfaked. We classified deepfake accounts as accounts that have posted at least 3 deepfake posts.

Deepfake text content detection works largely by examining 2 features, the perplexity and burstiness of a given text and generates a score. The perplexity score is generated by the randomness of the text, and burstiness measures the variation in perplexity. Chat GPT and other deepfake text creation AI typically generates text that lacks both randomness and variation. Because of this, AI text content detection algorithms work better with longer forms of text, usually 2-3 or more sentences. Therefore, we restricted the posts we analyzed to those of 100 characters or more.

A. Network Density

Our first method was to analyze how connected deep fake accounts are to each other. We believed this process would be helpful in determining the interconnectedness of deepfake accounts to each other. This process can be broken down into 4 parts. First, we found a popular twitter account that posts deepfake content. Next, we analyzed the twitter accounts following and followers list then determined if those accounts post deepfake content. Lastly, repeating step 2 over this set of accounts allowed us to create a network density graph for deepfake posting accounts.

Examination of the interconnectedness of the flagged accounts pose multiple benefits towards understanding and addressing the issue. Firstly, this investigation would allow for researchers to get a grasp on the size and scope of the problem at hand. Mapping together a network of accounts flagged for posting artificially synthesized text would permit for the understanding of several necessary questions pertaining to the quantification of the issue, including: the ratio of accounts that are involved as opposed to merely consuming the media, through what means are the flagged accounts are connected, as well as answering what types of content the accounts were similarly posting. Through answering these questions, researchers could potentially identify the bad actors who are using these new technologies to manipulate public opinion and inflict harm.

Part one of this method was done by manually searching through twitter to find a popular twitter account that regularly posts deepfake content. We defined a popular account as having over 3000 followers, and a deepfake account as one that posts at least twice a week with the content being 30% deepfake. For part two of this process we compiled a set of this accounts follower and following lists, because follower and following lists are public information this was done through web scraping the necessary data. Part three of this process was done by iterating over this list and determining which account was a deepfake account using the same constraints as part one excluding the popular account requirement. These accounts were iterated over starting with part 2 until we had a considerable amount of data.

B. User Engagement

The second method that we implemented measures user engagement over different deepfake accounts and analyzes their account growth. We believe this method would help us gain insight into the degree of what deepfake content is influencing public opinion. The first step was to compile a list of deepfake posting accounts. Next, we measured user interactions on their last 5 posts and recorded them along with their following account. Every 10 days, we repeat this process for each account. This allows us to compare account growth over time by user interaction. As of this report we are still recording data from these accounts.

The first step, finding deepfake accounts, was done in a similar way to that of the first method, where an API was used to determine if an account was a deepfake account. To measure user engagement on each account we took their last 5 posts and evaluated them on a point basis, comments being 5 points, retweets being 3, and likes being 1 point. As with follower and following lists, comments, retweets, and likes are public information accessible through web scraping. Each account's points were recorded into a text file along with their following count and account name. When recording each account's points every 10 days, we made sure to not double count posts that were counted by previous recordings.

C. Geographical Location

Now that the "Who" question had been thoroughly explored via the analysis of the previous two methods, it was time to turn our attention to a different approach in quantification: the "Where". The third and final method that was examined for the purpose of determining the best way to quantify deepfake content was exploring the geographical locations of the tweets themselves. What region of the world were these tweets emerging from and what kinds of audiences were their content directed at?

Taking the geographical location of the tweets into account, we would be able to gain valuable insights into the distribution of deepfake content and identify which regions or countries are being most impacted by the effects of artificially generated social media content. Being able to identify the patterns and trends in the emergence of deepfake content could assist in mitigating the potential harms they might cause through tailored awareness campaigns or educational resources to combat the spread of false information.

In practice, based upon the data collected from the tweets that our team had access to, there were ultimately two characteristics that we could examine that could help answer this "Where" question. Firstly, the location of accounts could be garnered from the location displayed on the account level of both the flagged accounts as well as their followers. These locations were stored in the database to be used to compare which regions of the world tended to follow and be more perceptible to artificially generated content.

As a second metric to further categorize the tweets into groupings that demonstrated which regions of the world the tweets were interacting with, the language that the tweets were tweeted in would be examined. This task was accomplished using the detectlanguage-python API client. Detectlanguage accepts the contents of the tweets as text, discerns what language the tweet is in out of the 164 total languages that it can choose between, and then returns the top language along with a confidence score. We were then able to attach this information alongside each tweet that arose from accounts flagged for deep fake content in our database to categorize the tweets into groupings of different languages to be compared on an interval of every 10 days.

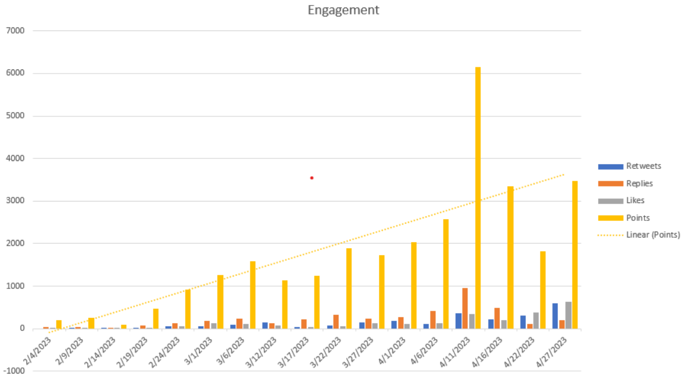

Fig. 1. Histogram of number of followers on deep fake accounts

III. Evaluation Metrics and Graphs

As previously stated, we devised three methods to measure the prevalence of deep fake content on social media platforms. The initial approach involved generating a density graph that depicted the connections between deep fake accounts and accounts we identified as genuine users. However, the resulting graph consisted of numerous unconnected nodes, yielding limited insights into the emergence of deep fake content. We conclude that a more substantial sample size is necessary to conduct meaningful analysis using this method. The second method is quantifying user engagement (followers,likes, comments, reposts) from accounts that post deepfakes and their growth.

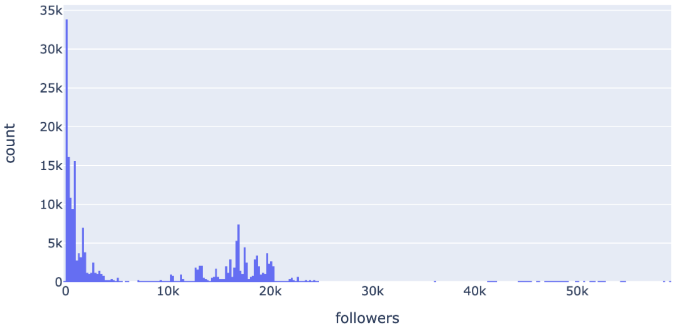

Fig. 2. Histogram of number of followers on deep fake accounts

Displayed above is a histogram representing the distribution of followers amongst the deep fake accounts within our dataset. Notably, two significant clusters are present: one comprising accounts with fewer than 6,000 followers and the other containing accounts with between 10,000 and 20,000 followers. This observation suggests that certain deep fake accounts have garnered substantial audience sizes. However, it is possible that some of these followers may be artificial bots created to boost engagement.

The third method we planned to quantity the emergence of deep fake content on social networks is by seeing the growth of this accounts segmented by region.

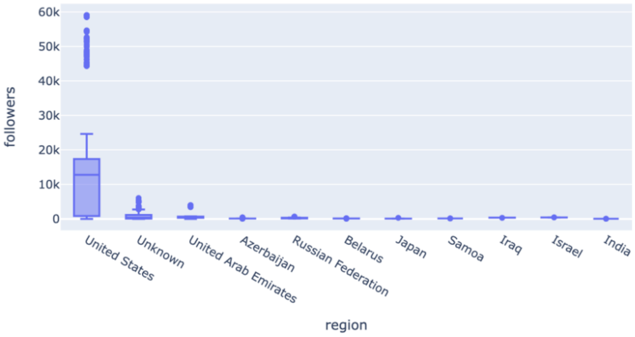

Fig. 3. Box Chart of number of followers on deep fake accounts by region

Shown above is another graph, partitioned by region, displaying a box chart of deep fake accounts and their corresponding number of followers. As observed, our data is heavily skewed towards the United States. Nonetheless, it is noteworthy that countries, such as Russia and its close ally Belarus, which has been accused of employing ”troll deep fake accounts” to provoke social unrest, are also represented in the chart. In addition, it is difficult to ascertain whether our sample of accounts is a truly representative reflection of the broader population of deep fake accounts. It remains unclear whether there is, in fact, a disproportionate number of deep fake accounts in the U.S. relative to other countries or if our data collection methodology may have data sampling bias of selecting U.S accounts. Finally, of all the accounts we scraped, 97.51% were identified as genuine users while 2.49% were identified as deep fakes.

IV. Conclusion

After collecting data with our three methods, it appears that user engagement is the best method we have found to quantify the emergence of deep fakes text on Twitter. This makes sense, because although we do not have access to Twitter's algorithm, growth on social media platforms is generally tied to the amount of engagement a user's content receives. The visualization of engagement with deep fake content with respect to time shows a growth suggesting a positively sloping trendline. This leads us to conclude that user engagement is an appropriate method to gauge the emergence of deep fake text on Twitter. As this technology becomes more pervasive and convincing, not just in text media but also in other forms of media such as image and video, it will be important for us as a society to become increasingly wary of the content that spreads online.

References

[1] Pu, Jiameng, et al. “Deepfake Videos in The Wild: Analysis and Detection.” Proceedings of the Web Conference 2021, 2021, https://doi.org/10.1145/3442381.3449978.

[2] Pu, Jiameng, et al. “Deepfake Text Detection: Limitations and Opportunities”, 2022, https://doi.org/10.48550/arXiv.2210.09421.

[3] D. Güera and E. J. Delp, ”Deepfake Video Detection Using Recurrent Neural Networks,” 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639163.

[4] Li, Yuezun, and Siwei Lyu. ”Exposing deepfake videos by detecting face warping artifacts.” arXiv preprint arXiv:1811.00656 (2018).

[5] Diakopoulos, Nicholas, and Deborah Johnson. ”Anticipating and addressing the ethical implications of deepfakes in the context of elections.” New Media & Society 23.7 (2021): 2072-2098.

[6] Matern, Falko, Christian Riess, and Marc Stamminger. ”Exploiting visual artifacts to expose deepfakes and face manipulations.” 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE, 2019.

[7] P. Korshunov and S. Marcel, ”Vulnerability assessment and detection of Deepfake videos,” 2019 International Conference on Biometrics (ICB), Crete, Greece, 2019, pp. 1-6, doi: 10.1109/ICB45273.2019.8987375.

[8] Zhao, Hanqing, et al. ”Multi-attentional deepfake detection.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[9] Lyu, Siwei. ”Deepfake detection: Current challenges and next steps.” 2020 IEEE international conference on multimedia & expo workshops (ICMEW). IEEE, 2020.

[10] Guarnera, Luca, Oliver Giudice, and Sebastiano Battiato. ”Deepfake detection by analyzing convolutional traces.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020.

[11] Rana, Md Shohel, et al. ”Deepfake detection: A systematic literature review.” IEEE Access (2022).

[12] AI content Detector: https://writer.com/ai-content-detector/

[13] Language detection API: https://detectlanguage.com/