Project Proposal - Quantifying the Emergence of Deepfakes

Kaushal Bhat
Carlos Field-Sierra
Anthony Bisgood
Chase Klapperich
Joseph Clancy

## Introduction

*Research Question*

The continued proliferation of deepfake technology and the content it produces has resulted in new challenges to social media networks, weaponizing information in a way that takes advantage of the online ecosystem. Deepfake content, or images and video that have been created via feeding a large dataset of existing media of a particular individual to a generative deep learning model to produce new, artificial-intelligence generated content that often mimics the appearance or speech of said person, has serious potential to effect harm on both a societal or personal level in the hands of a bad actor. In light of these concerns, it has become absolutely essential to understand how to best quantify the emergence of deepfake content on social media platforms.

To this end, the proposed research project will explore the following question: How can we best quantify the emergence of deepfake content on social media networks? This project makes use of a combination of analytical techniques in an attempt to arrive at a comprehensive understanding of how the online world is being impacted by this emerging deepfake technology.

Among the methods employed, this study will look at the network density of various accounts flagged for spreading artificially generated content to put into perspective the concentration of deepfake content within various social media circles, as well as conduct time series analysis to determine the rates of growth of both the posts and the content production over time. Sentiment analysis on the contents of peripheral posts will allow for classification of the content as either positive, negative, or neutral, aiding in determining the overall emotional impact that these posts have on those that are interacting with them. Additionally, timestamps and metadata of the examined social media posts will be used to aid in the identification of any patterns in the distribution of AI-generated media. The utilization of these techniques across different dimensions such as the region or location of the post, follower count and reach of the poster, and the age of the account, will bring to light both how and in what ways algorithmically generated media is emerging on social media networks.

*Motivation*

At the breakneck speed of development that artificial intelligence is currently undergoing, it is becoming increasingly apparent that the implications of such a powerful technology are fierce and far-reaching. Its usefulness in art and media has already been demonstrated, such as in the Star Wars film Rogue One involving a performance from a deepfaked young Carrie Fisher[1] or John F. Kennedy's voice reading a speech that he never got to make[2]. However, in the hands of an adversary these innovations turn from beneficial tools, aiding in bringing one's imagination to life, to weapons, capable of stoking social unrest and political polarization.

Through understanding and quantifying how deepfakes propagate across digital communities and onto the social media feeds of unsuspecting users, it becomes possible to assess the overall impact that they have on society as well as mitigate potential damages that result from their spread. Knowing who or what deepfake content is trying to imitate

provides a more thorough understanding of both the context and the message that is being pushed by the creator as well as their motivations. This information can then be utilized to critically evaluate the content and make informed decisions about the moderation of the material based on its credibility and potential impact, allowing social media companies to improve the detection and prevention of harmful deepfaked materials. Improving upon these standards for moderation aids in building and keeping authenticity and trust in the information present on a given site, simultaneously protecting reputations of those at risk and preventing misinformation to run rampant.

In addition to the foreseeable gain in security, quantification of the spread of deepfakes is critical to laying a solid foundation for the future of deepfakes both in the courtroom and in cybersecurity labs, as future research is conducted and their legality discussed. The policy surrounding artificially-generated media is still in its infancy and there is a growing need for policymakers to take action to address the potential consequences for journalism and public discourse in addition to the privacy and security concerns. Through understanding where deepfakes are coming from and the messages that they carry with them, policymakers can create regulations that manage the use of deepfake material, while also ensuring that freedom of speech and privacy rights stay protected. In the laboratory, this information will be critical to help future researchers identify patterns and trends in deepfake use and abuse, providing a basis for developing novel techniques and tools for deepfake detection. In conclusion, the quantification of deepfake content on social media networks is essential to understanding and addressing their potential consequences, regardless of the field.

*Related Work*

Related work in in the field includes papers such as “Deepfake Text Detection: Limitations and Opportunities””[3] and “Deepfake Videos in the Wild: Analysis and Detection”[4], both of which analyze the current state of deepfake detection technologies and propose new approaches to detecting deepfake content in the wild. In “Deepfake Text Detection”, researchers explore current text deepfake detection by creating several novel ways of circumventing deepfake text detection technology and propose a new way of text detection by “tapping into the semantic information in the text content” [3]. In “Deepfake Videos in the Wild”, which explores how deepfake detection defenses do against real-world deepfakes, as compared to pre existing datasets. To do this researchers collected “the largest dataset of deepfake videos in the wild " [4] and analyzed the popularity and growth of that deepfake content. Researchers then evaluated existing deepfake detection algorithms against their new dataset, and explored ways to improve defenses.

*Our Contribution*

Our contributions to the field of quantifying the emergence of deepfake content on social media will give researchers new insights on how deepfake content continues to evolve. In particular, research papers [3,4] explore how well current deepfake defenses can identify deepfake content in self created datasets versus existing deepfakes on social media platforms. Therefore creating better ways to quantify the emergence of deepfakes, (ie growth, network density, etc.) will allow researchers to test and develop new deepfake detection algorithms on the best, real world deepfakes.

## Methodology

Overall there are three key stages: data collection, data organization, and data analysis for insights. Each step presents various options and trade-offs that must be considered and made as a group. Currently, we have not come to a decisive conclusion on what paths to make, therefore, I will detail each step separately with a paragraph for each below.

First we must collect the data on social networks. This task brings up a lot of questions, such as which social networks to include, what time frame to consider, and how much data to collect. The answers to these questions hinge on our chosen method for obtaining the data. We have two options to consider: scraping the data ourselves from the internet or relying on an existing data set. Scraping the data ourselves offers greater control over the data we obtain and ensures it is up-to-date. However, this method may yield a smaller amount of data compared to using an existing data set and may have a limited time range since we are starting from scratch. The major drawback of this approach is the requirement of a large sample size for statistical significance. It is uncertain if we can obtain a sample of sufficient size to meet this requirement. In comparison, an existing data set would provide a larger volume of data and a longer time range, but it comes with less control over the collected data, potential outdated information, and uncertainty over its accuracy. Lastly, if we opt to collect the data ourselves, ideally we would need to run a bot in the cloud and store the extensive data in a platform such as BigQuery or a similar solution.

The next step is organizing the data. It is a cliche that in data science most of the time is spent cleaning the data. Although time consuming this can be done trivially with pandas in python or a similar library. In addition, we need to label the data as deep fake or not deep fake. Doing it manually would be out of scope for us, so the most practical solution is to use existing APIs for this purpose. Companies that provide deep fake detection services include Hive, Reality Defender, and others.

Finally, we need to analyze the data. This can be achieved using Pandas and plotting libraries in Python such as Plotly to manipulate the data and gain insights. Additionally, statistical packages like scikit-learn can be used. This will allow us to answer the research question by exploring the trends & data we collected around deep fakes in social media.

**Evaluation Plan**

We intend to gain better insight into the growth of deepfakes on social media, as well as the best methods of measuring this growth. To do this, we will interpret the data we will collect about the social media deepfake situation from a few different perspectives. We will analyze the rate of growth of deepfake-dedicated accounts and posts in terms of quantity. We will analyze network density to see how social media deepfake interaction is evolving and if it can be useful for detection. We will visualize our data and analysis neatly in the form of timelines and network graphs.

Two main methods that we will evaluate are deepfake growth and its association with regions and how users interact with deepfakes. Finding a correlation between growth of deepfakes and regions they originate from in the metadata would be very beneficial to detection algorithms as well as predicting how deepfakes may be used in the next few years. Network density analysis results will also be evaluated. Deepfake-dedicated accounts may be seeing greater success, such as receiving and maintaining a higher level of engagement from users. If deepfake accounts are increasingly interacting with each other, a higher quantity of posts may

not indicate a growth of success; however, it could help detection efforts by association. Analyzing the network density will allow us to answer this important question that could be beneficial to detecting deepfake accounts.

## Execution Plan

The first major milestone for this project is having this proposal finished and thus having a tentative plan to start executing on. Our next milestone will be accessing data. Following this, having our data organized and ready to begin analysis is an important step. The next milestone is having analyzed the data, and the final milestone will be having compiled our report with findings based on our analysis of the data.

These milestones are not set in stone; for example, during the analysis stage, we may realize that we need some more data or that it would help to have our data organized differently; in cases such as these, some milestones will have to change. Because of this, we intend to be flexible and leave buffer time towards the later stages in case a need is realized for us to backtrack. That being said, we are planning to have accessed our data in the next 2 weeks, having organized it by the end of February. We want to have our data organized by the midpoint of March, and spend the remainder of March on the analysis. By the beginning of April, we want to begin writing about our findings based on our analysis, leaving us time to go back to previous stages if necessary, but also recognizing that with other final projects and finals looming, individuals' time available for this project may become limited.

Since we do not yet have a collective and concrete idea of each individual's strengths and time available over the course of the semester, we have not assigned concrete roles for the entirety of the project. But much like we did for this Proposal, we plan on meeting often and early to decide on these dynamically depending on the needs of each stage of the project.

## References


[1] Abrams, J. (Director). (2019). Star wars: Episode IX – The rise of Skywalker [Film]. Walt Disney Pictures; Lucasfilm; Bad Robot.

[2] Corcoran, K. (2018, March 16). *The speech JFK never got to give on the day of his assassination has been recreated with voice technology*. Business Insider. Retrieved from https://www.businessinsider.com/jfk-speech-from-day-he-died-recreated-with-voice-tech-2018-3

[3] 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.

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