Samsara Counts & Dominique Dalanni

Using Deep Learning to Improve Eating Disorder Treatment

Overview

Eating disorders (ED) are serious medical conditions that can have disastrous effects on an individual's mental and physical health. They are pervasive, and do not discriminate based on race, religion, gender, or socioeconomic status. ED are often a lifelong struggle, with approximately ⅔ of patients never achieving a full and sustained remission. They are the product, in part, of increased societal pressures to fit "the thin ideal," and exposure to this media can be triggering to people with ED as well as those at risk for developing them. Social media platforms are especially rife with these triggers—individuals with ED have created communities where they support one another in the dangerous pursuit of being "thin enough." These websites teach readers how to act on and hide their ED, putting them at risk for severe physical and mental health complications, including death. Therefore, triggering content online poses a serious risk to social media users with ED and those at risk for developing them. Similarly, it is essential that clinicians and family members be able to identify websites containing images that are associated with the promotion of ED to prevent accidental or intentional exposure to these triggers. However, it is challenging for caretakers to find and stay up-to-date with ED communities and content online. This research aims to automatically detect such triggering material, with the ultimate goal of designing tools to inform clinicians and support patients in their recovery. The main products of this work are a convolutional neural network that identifies images of ED and two novel software tools built from it that assess websites for ED content. These tools would enable clinicians, family members, and those suffering from ED to understand and identify sources of ED content to improve treatment for ED patients.

Intellectual Merit

While other researchers have used machine learning to automatically detect pro-ED content, no one has made a *robust* deep learning (DL) classifier that accurately detects images of ED. Most importantly, there are no user-centric tools using DL to improve health outcomes for ED patients, so both tools are one-of-a-kind. With the first tool, clinicians can access a newsfeed depicting the latest trends in pro-ED communities online. The tool also explains which visual features distinguish ED images from other similar image categories. This feature uses a class activation mapping, a DL interpretability technique that has never been implemented in this context. The second tool is for patients in recovery: a browser extension that uses the classifier and text-based machine learning methods to filter triggering webpages. Finally, related work has not incorporated measures of fairness and human diversity into classification; thus, other classifiers likely include biases towards demographic groups underrepresented in the classifier's training dataset. This project combats this bias in two ways. First, the classifier imposes algorithmic group fairness guarantees such as equal false positive/negative rates to improve classification fairness across images with subjects from different demographics. Second, during training the algorithm, images from underrepresented demographic groups are strategically sub-sampled to improve classification accuracy.

Broader/Commercial Impact

This project has the potential to improve health outcomes for ED patients in practice as well as streamline the process of developing accurate, unbiased image classifiers. The ED classifier would be the first of its kind, demonstrating that it is possible to use DL to automatically detect images of ED with high accuracy. The two software tools are also the first of their kind on the market. With the first, ED clinicians could understand the latest trends in pro-ED online media in order to better treat their patients. With the second, ED patients could have a better recovery by seeing less triggering content without impeding their internet use. Finally, this project would contribute to the body of work using deep learning for recognition of objects in high-level categories, as well as incorporating fairness guarantees into classification accurately.

Elevator Pitch

According to the National Association of Anorexia Nervosa and Associated Disorders, at least thirty million people in the United States suffer from an eating disorder[1] (ED). ED are complex medical conditions with the highest mortality rate of any mental disorder[2]. While ED are generally associated with white women, they affect people of all genders, races, and walks of life. Treatment of ED requires long-term rehabilitation, a strong support system, and sustained behavioral changes, similar to drug dependency rehabilitation. One major setback to recovery from an ED is exposure to triggering images or content that can bring back disordered thoughts and behaviors. This content poses a serious threat to patients, yet can be found anywhere, particularly online and on social media—yet there exists no way to avoid it without avoiding certain platforms altogether, perhaps even the internet, which a substantial handicap to that person's daily life.

On the other side, it is challenging for clinicians to find such material without a familiarity with the host platform, not to mention understanding trends in posts. This, too, is a serious obstacle, because it is critical for medical professionals who treat patients ED to stay informed on what risks exist outside of treatment. Similarly, family members and caretakers of people with ED would benefit from stay up-to-date with trends and challenges their loved ones might encounter.

The proposed two software solutions that assess websites for ED content with distinctly different purposes: one for diagnostics, the other for filtering out triggering material. The first is intended for clinicians and caretakers, and the second is for patients. Both tools would enable users to understand and identify sources of ED content with the goal of improving treatment and recovery outcomes for ED patients. Currently, there exists no tool to filter this content available on the market nor software designed for ED clinicians or patients. Therefore, these tools would be the first of their kind, in addition to solving a challenging problem at the forefront of deep learning technology.

Behind the scenes, the proposed tools detect and process triggering images using a deep learning classifier—a convolutional neural network—trained for this specific task. The classifier is optimized for algorithmic fairness as well as performance, achieved by programmatically incorporating diverse data in the training set to obtain a dataset that is more representative of the demographic range of triggering content. The classifier was designed by combining concepts in fairness in machine learning and computer vision during dataset preprocessing and algorithm training to maximize its performance in a challenging problem domain with unbalanced data classes. Unbalanced can mean different numbers of examples in different categories, but can also mean that within a category there are features that are under-represented. In the context of diagnostic imagery, ensuring that classifiers work on unusual cases is important. This product embodies a novel, cutting-edge approach to improve classification performance and fairness within the context of accurately classifying images of eating disorders. Hence, this project has both the life-altering potential to improve health

[1] http://www.anad.org/education-and-awareness/about-eating-disorders/eating-disorders-statistics/

[2] http://www.anad.org/education-and-awareness/about-eating-disorders/eating-disorders-statistics/

outcomes for ED patients in practice and streamline the process of developing accurate, unbiased deep learning classifiers.

The Commercial Opportunity

The main market for this product will be the various facets within healthcare industry concerned with diagnosing, treating, and preventing the recurrence of eating disorders. However, the thirty-million individuals in the United States who are currently suffering from an eating disorder provide a large market for the patient browser extension due to its ability to omit any triggering content, which can be a vital asset for patients in recovery. Because of this, there will definitely be a market for the product for patients going through post-hospitalization and outpatient recovery. Hospitals with eating disorder rehabilitation units will also benefit greatly from having this software product for clients, patients, and the individuals who make up a patient's support system. Small clinics focusing on eating disorder rehabilitation will also be able to utilize this product in their eating disorder rehabilitation and recovery programs. Mental health facilities can benefit from the product, primarily through the use of the eating disorder recognition software toolset. These tools will also be particularly useful for emergency short term care facilities, and long term treatment centers as well. College campuses and their research and medical programs provide yet another market. Students may benefit from this product, and may be able to research its long-term effects on the prevention of eating disorder relapse. Collegial medical programs provide a market as well due to their desire to train students to diagnose, treat, and prevent illness.

This product has a two-tiered business model which focuses on healthcare providers and patient services. The primary customers for the product will be caretakers, clinicians, therapists, and the individuals who make up a patient's support system. Patients participating in eating disorder rehabilitation and recovery will be the other main customer group.

While there are eating disorder-specific recovery applications (apps) for mobile phones, all of them focus on either connecting users with other people in recovery or documenting the user's recovery journey. Some of these apps also allow users to connect with their treatment team, including therapists None have the goal of helping users avoid potentially triggering content on other websites and none are designed for use on a personal computer.

One of the most promising aspects of this product is the fact that there is currently nothing else on the market. This creates the potential for limitless possibilities and growth in a field, while bringing an extremely promising and helpful product to the market.

The primary risks of the product involve potential violations of international data privacy laws (and accusations thereof). For example, if any of the product's users are citizens of European Union, then the product must adhere to the standards of the General Data Protection Regulation (GDPR) when processing personal data. Under the definition of GDPR, the software is a data processor of personal data, and there are eight fundamental data subject rights data processors must comply with. To handle this, all software was designed to be GDPR compliant (regardless of where users come from): no publicly-available data is ever stored and it is only collected in real time from public websites. Hence, if a web page is removed from the internet but a user tries to analyze that page, they cannot. Additionally, both the software and the classifier's training and test datasets are closed-source to ensure privacy and help prevent malicious use of the software.

Finally, the other potential risk of deploying the product is users misusing it with ill intent. This possibility exists for both the patient and clinician tools, though it is more of a concern with the patient tool. For that particular tool, there is a chance that users could use the browser extension to gain information about which websites are focused on eating disorders. They could then use this information to spread misinformation and/or slander an individual or organization based on how the software classifies their website. Though this cannot be entirely avoided, to address this, there will be a terms of use contract for the software. The software will also have a feature where users in violation of the terms of use can be revoked access to the classifier (thus rendering the browser extension useless). With regards to the clinician tool, there is some potential for privacy violations in the form of ED analysis of a particular website if an adversary were to use it. To prevent any misuse, the clinician tool has extremely restricted use: only people with a proof of employment as a medical professional or eating disorder clinician can gain access to it.

In the United States, this product has a significant revenue potential from long term health care contracts and the continued support, treatment, and recovery plans associated with people suffering from, or who have suffered from, an eating disorder. If one were to assume that the product positively affected the thirty million individuals in the United States who have suffered from an eating disorder, and each person saved just forty dollars per year in healthcare fees, that would come to about 1.2 billion dollars saved annually. While these numbers have not been tested or proven, it is safe to assume that this product does in fact hold a great potential for revenue in the form of health care provider contracts in addition to its potential to save

Healthcare facilities and insurers will be more likely to invest in contracts for using this product due to its long term saving potential of healthcare costs associated with long term eating disorders. Patients who are successful in eating disorder recovery efforts will cost the healthcare system and healthcare insurance industry less across their lifespan. It will be cheaper for healthcare facilities and insurers to invest in the use of the eating disorder analysis software suite, rather paying for the long term costs associated with unsuccessful eating disorder treatment, and the underlying long term health issues eating disorders cause.

Societal and Global Impact

There is substantial social and commercial opportunity for the product to provide a method for easing recovery from a major mental health issue. In term, the product can also help lessen and possibly prevent long term associated health issues associated with eating disorders, such as heart arrhythmia, anemia, and other life-threatening conditions. Furthermore, this product gives an opportunity to help shift the affected population towards healthier habits. In turn, as users are able to recovery, there will be an increase in healthy individuals able to contribute more to society.

The eating disorder analysis software suite can potentially save billions in healthcare costs associated with the long term health effects from eating disorders. Through this product, patients in recovery gain more tools for success and relapse prevention, which will in turn lead to an increase in their long term health. Healthier people can work and be better economic consumers, as they are better equipped to contribute to society, and this product definitely provides one way to approach this goal.

The software suite aims to help clinicians, family members, and those suffering from eating disorders to understand and identify sources of eating disorder content to improve treatment for patients. Thus, this product could impact people from all demographic groups across the United States and potentially the world.

First and foremost, this project benefits people with eating disorders all over the world. The plan to release the patient tool—a web browser extension to filter out triggering eating disorder-related content—is to offer it for free. Hence, this project is accessible to and designed to benefit people from all backgrounds and walks of life. This software could be installed in public computers at treatment centers and wards hospitals that treat people with eating disorders, so patients in intensive inpatient or outpatient programs can still use the internet without seeing triggering content. This tool would also be useful for parents, particularly parents of young girls (who are especially vulnerable to developing eating disorders), who want to make sure their children can browse the internet without encountering pro-eating disorder communities.

From the caretaker perspective, this project benefits both medical professionals and the family and friends of people with eating disorders. The clinician software enables users to see the latest trends of pro-eating disorder communities online, in order to understand what their patients may face. It also includes a diagnostic tool to instantly assess how eating disorder-focused a particular website is, which would be useful for quickly analyzing blogs of patients, to get a sense of what kind of content users have posted over time.

For friends and family members of people with eating disorders, this tool could give them insight into the world of people with eating disorders on the internet. Furthermore, the discovery feature for online eating disorder communities could help them empathize with and visualize the challenges their loved one is facing. The diagnostic feature could give them context on how eating disorder-focused their loved one's blog is. Finally, for online platforms where eating disorder communities are prevalent, the discovery tool could help them deliver resources to users who engage in this behavior, such as links to support hotlines and websites.

While it is important to note that there are environmental issues associated with building the hardware used in computing devices, the product itself does not have any associated environmental issues to date as it is purely software-based. Since the clinician toolset contains images and content from pro-eating disorder communities, it would not be appropriate for children or vulnerable populations (such as patients in recovery). However, children may use the patient tool as it only filters out eating disorder-specific content. Major regulation for this product is not necessary, as it already falls under the governance of global data privacy laws such as the General Data Protection Regulation in the European Union. However, restrictions will be put in place by the designers as to who has access to the clinician toolset to prevent potential misuse, such as requiring proof of employment as a clinician at an eating disorder clinic to gain access.

Since the product filters internet content, it is possible someone could use the product in an unethical manner for censorship purposes. To subdue this risk, the classifier will remain closed source to ensure it is used only in the manner in which it was designed for: as a tool for eating disorder treatment and recovery. Additionally, by restricting who has access to the clinician toolset, unethical use will be prevented.

With an estimated 70 million people with eating disorders worldwide, they are clearly already a serious global health issue. The proposed software suite can help alleviate some of the struggles of eating disorder recovery for people that use the internet, ideally making it easier to resume a healthy life and avoid triggering images.

Another potential global impact of the product is its effect on pro-eating disorder online communities, which exist in many diverse parts of the internet and involve people from all over the globe. Ideally, the product would not greatly affect such communities and users other than providing a resource for members to use in recovery. It is possible, however, that the reach of the product could be misconstrued as a censorship and moderation device and community members could react negatively. For example, similar to what occurred when online platforms attempted to moderate pro-eating disorder forums based on the keywords posters use, communities could withdraw and become more private and hard to detect. Ideally, this will not be an issue, as the product will be exclusively advertised as a tool for recovery, not a censorship or moderation device.

Technical Discussion and R&D Plan

Innovation Description

Eating disorders (ED) are serious medical conditions that can have disastrous effects on an individual's mental and physical health. They are pervasive, and do not discriminate based on race, religion, gender, or socioeconomic status. One issue that makes ED particularly difficult to treat is a patient's exposure to triggering content online. While there are eating disorder-specific recovery applications (apps) for mobile phones, all of them focus on either connecting users with other people in recovery or documenting the user's recovery journey. Some of these apps also allow users to connect with their treatment team, including therapists as well.

However, none of the applications currently available to date are designed to help users avoid potentially triggering content on other websites and none are designed for use on a personal computer. The proposed technology provides a solution to this dilemma, and involves a two-tiered business model which focuses on healthcare providers and patient services. The first tool, explainED, will provide ED clinicians with a newsfeed style overview of ED trends, graphs, and patient information. The second tool, *filterED*, will provide ED patients with a method to filter out triggering content while browsing the internet.

The primary technical challenge to be addressed in Phase 1 is adding a diversity function feature to the classifier. Currently, the deep learning classifier used for this project does not adequately recognize diversity features, which results in biased classification results. By optimizing the diversity function and picking the most diverse images possible at training time, the classifier will be trained on the most diverse data with the goal of performing accurately on diverse classes in the testing phase. Therefore, the filtering of ED related content will be improved because the classifier will have a higher success rate of filtering out diverse ED content.

One potential risk of deploying the product is users misusing it with ill intent. This possibility exists for both *explainED* and *filterED*, though it is more of a concern with *filterED*. For that particular tool, there is a chance that users could use the browser extension to gain information about which websites are focused on eating disorders. They could then use this information to spread misinformation and/or slander an individual or organization based on how the software classifies their website. Though this cannot be entirely avoided, to address this, there will be a terms of use contract for the software. During Phase 1, a software feature will be added that revokes access for users in violation of the terms of use to the classifier, thus rendering the browser extension useless. With regards to *explainED*, there is potential for privacy violations in the form of results of the analysis of a particular website being misused. To prevent any misuse, *explainED* will have extremely restricted use: only people with a proof of employment as a medical professional or eating disorder clinician can gain access to it. This will be guaranteed through a vetting process.

Phase 1 Objectives

For Phase 1, both the User Database and Research Database must be fully operational, and all images and data must be migrated into the Research database. A considerable portion of the frontend framework for both *explainED* and *filterED* must be implemented, and must communicate with the database. The patient tool *filterED* is a Google Chrome browser extension, and will be implemented to run in the browser as an add-on to other websites that runs in the browser. It will use the webrequest Google Chrome API for HTTP communication and interacting with web pages.

The clinician tool web application, *explainED*, uses HTML and Javascript for both frontend web development and communication with the server. *explainED* is a dynamic web application that displays trends in pro-eating disorder content online, provides an analysis of a given social media profile or website, and visualizes trends in ED data. The local copy of the classifier, hosted on the webserver performs quick image analysis from the two client-side applications (to avoid the latency of communicating with the other server). To display trends in the data, *explainED* makes requests to the research server for specific data including ED trends, patient trends, and data to be used for graph visualization. Requests will be made to the web server's user database to access and store user data. The frontend of *explainED* is implemented using Bootstrap templates with HTML and Javascript. Graph visualization algorithms must be applied to improve data visualization, and will be displayed on the *explainED* newsfeed. Graph visualization will be dependent on HTML and Javascript, and functions through the use of Python's *PapaParse*, *C3*, and *D3* libraries. All analysis and processing of the data for visualization is implemented in Python.

Users for both *explainED* and *filterED* can register for an account. For *explainED*, clinician consumers will have the ability to create an account with a valid email address and clinician/healthcare identification number. Clinician consumers must be verified by a human expert as certified healthcare professionals before their account is activated. Once their information has been verified, they will receive an email message that contains a link to activate their account. Once the link has been clicked, the account is considered active, and they can log in on the website. For *filterED*, patients can create an account with a valid email address and password. Once a registration request has been sent, a requester will receive an email message that contains a link to activate their account. Once the link has been clicked, the account is considered active. This implementation will be completed during Phase 1 as well.

The main backend functionality for both software applications is the deep learning classifier, implemented using the PyTorch library in Python. We used hashtags to identify training content for our classifier; we gathered images from social media platforms to train a classifier that can detect pro-ED content. Hashtags are user-defined and thus provide us with a means of access to images that are representative of the pro-ana community as defined by its members. Since #proana is a known identifier of a strongly pro-ED community, we used this hashtag as our starting point. We used a standard Convolutional Neural Network as the basis for our classifier.

The images in our final dataset came from Twitter, Tumblr, and Flickr, collected from over a period of six weeks. We removed duplicate images, but in all other respects, the dataset was unedited. Our training data was chosen to compare pro-ana content with other content similar in demographics and photographic style.

The diversity function must also be implemented with the deep learning classifier during Phase 1. The diversity capabilities of the classifier will provide the software with the potential to improve health outcomes for ED patients in practice as well as streamline the process of developing accurate, unbiased image classifiers. The refined model of the ED classifier would be the first of its kind, demonstrating that it is possible to automatically detect a diverse array of ED images with high accuracy. In addition, the classifier will have at least 85% overall accuracy and a comparably high recall score, in addition to demonstrably fair performance across demographic groups. The training set will have grown by at least a factor of 2 to ensure robust performance.

To get the product to market, the team must meet the following key technical milestones. The User and Research databases must be implemented, and the old and new data must be integrated into the database. *ExplainED* will be a functionally hosted website with the following capabilities: graph visualization that display trends in data, an analysis tool that provides classification results for an input website, a personal page per user with saved results. The *explainED* frontend template, including graph visualization, must be ready for integration during the next phase of the project. This includes the basic functionality of *filterED* as a working Chrome extension that communicates with the classifier and blocks pro-ED images and text. All registration and login pages will be created and ready for database integration as well.

The product has a high chance of commercial successes, as the main market for this product is a variety of entities within the healthcare industry concerned with diagnosing, treating, and preventing the recurrence of eating disorders. In addition, the thirty-million individuals in the United States who are currently suffering from an eating disorder provide a large market for the patient browser extension due to its ability to block triggering content, which can be a vital asset for patients in recovery. Hospitals with eating disorder rehabilitation units will also benefit greatly from having this software product for clients, patients, and the individuals who make up a patient’s support system. Most importantly, there is no existing competition on the market for this product, and the product is hosted on google chrome and the internet, which are both very popular and accessible to clients.

The following is the timeline of our R&D progress along with accompanying descriptions of each deadline and the expectations for the functionality of each.

15 December 2018

filterED, the chrome extension, successfully filters out pro-ED words from a predefined list when activated in the user’s browser.

25 December 2018

Long term scraping scripts will be deployed on the Research database to gather more images for training.

1 January 2019

Acquire over 50,000 new images for training. Then we will retrain the classifier with those images and compare performance. All scripts implementing automated retraining and graph generation will be completed. The webapp, *explainED*, will be set up on Amazon AWS.

10 January 2019

The database implementation and backend server settings will be complete. The server effectively communicates with the database. All past data will be successfully integrated with the database.

20 January 2019

Complete front end login authentication for *explainED* (the webapp). Complete frontend login authentication for *filterED* (the chrome extension). The frontend of the webapp effectively communicates with the server and the database by passing a username and password pair to the b, the database receives that pair and checks if they exist in the database. The server passes a 2 if they do, 1 if the username is right but the password is wrong, or 0 if they don't. The chrome extension then receives that data and either moves the user on to the next screen or displays an error message.

1 February 2019

All the existing images will be labeled with diverse features for retraining. There will have been at least 2 labeling sessions for the dataset. The existing frontend source for *explainED* will be integrated with the webserver and backend server. When someone visits the website, the login page will be there.

10 February 2019

The frontend login authentication for *filterED* (the chrome extension) will be complete. The extension effectively communicates with the webserver by passing a username and password pair to the server, the webserver receives that pair and checks the database if they exist. The server returns 2 if they do, 1 if the username is right but the password is wrong, or 0 if they don't. The chrome extension then receives that data and either moves the user on to the next screen or displays an error message.

`

20 February 2019

Diversity function incorporated into training the classifier. When given the command to retrain the classifier, the server will compute SURF features for the new images, run the images through the diversity function, and choose the most diverse set of images to serve as the training set.

2 March 2019

The newsfeed on the clinician side (*explainED*) will display static images and content from the database. *filterED* will successfully communicate back and forth with the webserver, sending a list of image URLS and receiving a list of classification results for each.

10 March 2019

Clinician users of *explainED* can add their patients to the app and link their patients' social media. *filterED* will communicate with the classifier and block images that are detected to be pro-ED.

20 March 2019

Clinician users of *explainED* can input a website and get an analysis of that website's content. They can then either choose to save those results and website to one of their patient profiles, or save them in their files.

1 April 2019

Both products are fully functional as according to the design document.