Recognizing Images of Eating Disorders in Social Media

S. N. Counts[1*] , L. Alkulaib[1*], J-L. Manning[1], R. Pless[1], J. Harnett[1], H. Xuan[1], R. Begtrup[2], D. A. Broniatowski[3]

Department of Computer Science[1]
Children's National Health System, Washington, D.C.[2]
Department of Systems Engineering[3]
countss, lalkulaib@gwu.edu

I. INTRODUCTION

Eating disorders (ED) are pervasive and do not discriminate based on race, religion, gender, or socioeconomic status. Comorbidities include anxiety, depression, substance abuse, self-injurious behaviors, and history of trauma. ED are often a lifelong struggle, with approximately ⅔ of patients never achieving a full and sustained remission.

ED are the product, in part, of increased societal pressures to fit "the thin ideal." These pressures come in the form of repeated advertisements on various media platforms, messages from the diet and exercise industries, fashion industry "norms," etc. Individuals who suffer from ED may have experienced trauma; the ED can provide a sense of control over these factors, albeit an invalid one.

Exposure to media expressing "the thin ideal" can be triggering to individuals with ED as well as those at risk for developing them. Social media platforms are especially rife with these triggers. Concurrent with the rise of social media, individuals with ED have created communities [1] in which they support one another in the dangerous pursuit of this illness' elusive goal: to be "thin enough." Websites promoting anorexia (pro-ana) and bulimia (pro-mia) as lifestyle choices valorize acting on ED symptoms. Such sites teach those suffering or at risk from ED how to act on and hide the illness, and support them in doing so, putting them at risk for severe physical and mental health complications, including death.

The impact of images in this community far exceeds that of other communities surrounding physical and mental health issues. Therefore, it is essential that clinicians and family members be able to identify websites containing images that are associated with the promotion of anorexia and bulimia to prevent accidental or intentional exposure to these triggers. This research aims to automatically detect such triggering material, with the ultimate goal of designing parental and clinical controls.

We report on a proof of concept, machine learning approach to identify pro-ana content, trained on example data from online social media searches. These proof of concept results suggest that it is feasible to automatically detect social media sources with triggering material, informing the creation of tools that can assist clinicians and family members to improve health outcomes.

II. METHODS

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 (Fig. 1), in the following categories:

- "pro-ana": 16,000 images from several Tumblr blogs including *best-thinspo*, *thinniest*, and *wanna-be-skinnyminnie*
- "selfie": 4,500 Tumblr images tagged "selfie"
- "ootd": 7,000 Tumblr images tagged "ootd" (outfit of the day)
- "Greek": 5,000 images from Tumblrs of Greek-letter college organizations.

SAMPLE CLASSIFIER TRAINING DATA
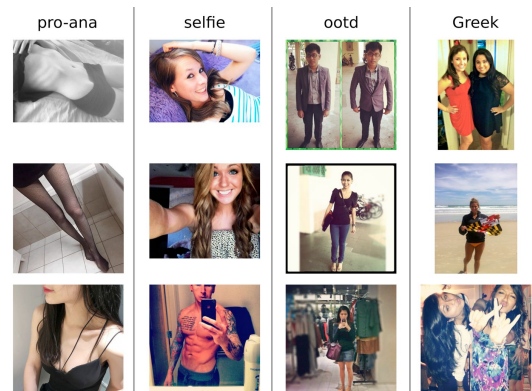
We randomly chose 4740 (15%) of these images to serve as test data and trained the Resnet Deep Learning neural

network [2] to classify the remaining training images into these categories.

III. RESULTS

On test data, our first iteration of data collection and training gives 78% classification accuracy—a significant improvement over chance (25%). To explore a possible application, we identify ten additional Tumblr accounts, five that we judged to have high pro-ana content, four blogs without pro-ana content, and one fitness inspiration (fitspo) blog that we found contained a mix of content. Figure 2 shows the percentage of images classified as pro-ana in each blog:

```
Blog type       Title                   %pro-ana
not pro-ana     abelmvada                  7
not pro-ana     roommysocks               21
not pro-ana     mathematicalmemer :        4
not pro-ana     satoshikurosaki           10
not pro-ana     traitspourtraits           9

pro-ana         oh2be-skinny              73
pro-ana         thinninglittle            88
pro-ana         think-skinny-th0ughts     89
pro-ana         oh2beskinny               83
fitspo          veganpilatesangel         53
```

Fig. 2. Table of classification results

Upon refining our dataset, we achieved nearly 81% accuracy after training. The pro-ana dataset, the positive category, included 10,393 images tagged with #proana. The negative category comprised of 31,197 images split evenly among images tagged with #selfie, #ootd, and #sorority. The data split was 80% of images in each class in the training set, with 20% withheld for the validation set.

IV. STRESS TESTING THE CLASSIFIER

A. Designing a Stress Test

To assess the robustness of our classifier, we examined its performance on images with subjects and demographics underrepresented in our training set. Thus, without changing our training data, we designed a stress test, with naturalistic examples from a variety of websites with content we thought the classifier might fail on (Fig. 3). In the pro-ana (PA) set, one site contained images of men; the rest were similar in content and demographic style to our pro-ana training set. For the non-pro-ana (NPA) set, we selected sites that were likely to show underweight people, a proliferation of exposed skin, or people of color. Our goal was assessing how the classifier performed with subjects of different races and genders as well as images with similar features. For example, noting the commonness of skin and erotic poses in the pro-ana image set, we included a portfolio of a boudoir photographer.
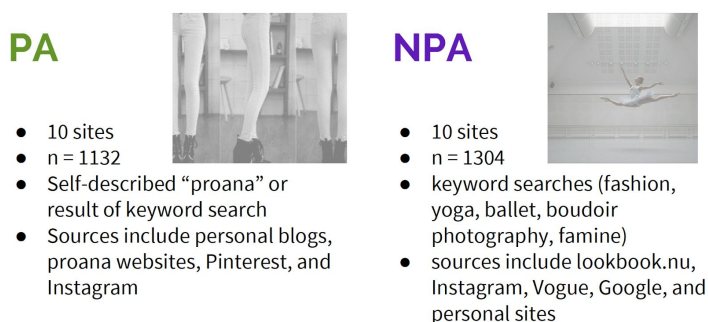

PA
- 10 sites
- n = 1132
- Self-described "proana" or result of keyword search
- Sources include personal blogs, proana websites, Pinterest, and Instagram

NPA
- 10 sites
- n = 1304
- keyword searches (fashion, yoga, ballet, boudoir photography, famine)
- sources include lookbook.nu, Instagram, Vogue, Google, and personal sites

Fig. 3. Description of stress test image sources and quantities

B. Stress Test Results

Figure 4 shows shows the percentage of images classified as pro-ana from each site in the stress test:

PA test set

i =	1	2	3	4	5	6	7	8	9	10
	100%	85%	100%	91%	91%	96%	99%	100%	100%	91%

NPA test set

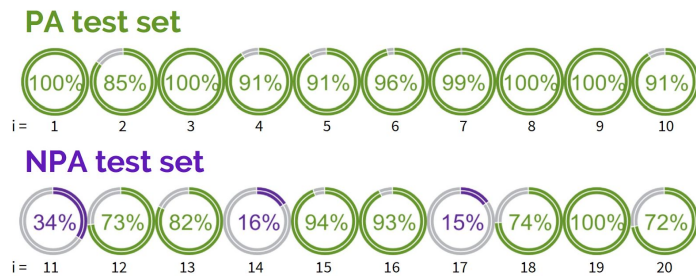| i =	11	12	13	14	15	16	17	18	19	20
	34%	73%	82%	16%	94%	93%	15%	74%	100%	72%

Fig. 4. stress test classification results from 20 unique sources

Despite our hypothesis that roughly half of the image types in the pro-ana test set would be difficult to classify, the results are encouraging. However, we do have a problem with pro-ana false positives. One potential explanation for this could be class overlap and noisiness of the dataset: for example, a pro-ana selfie is still a selfie.

VI. CONCLUSION AND FUTURE RESEARCH

We outlined our design and implementation of a machine learning classifier able to detect pro-ana images with 81% accuracy. In the future, we aim to improve the classifier's training dataset by gathering more images from each category.

Most importantly, since the classifier successfully identifies pro-ana images, we are using it to make a web application that assesses how pro-ana a social media user's content is. The tool, designed for clinicians, would allow them to enter a social media username and would then give an analysis of that user's online presence. The analysis would retrieve the account's history and classify how pro-ana its content is. The tool would also display a hashtag similarity map to show trending hashtags closely related to #proana.

VII. REFERENCES

[1]   A. Oksanen, D. Garcia, P. Räsänen, "Proanorexia communities on social media", *Pediatrics*, 2016.

[2]   K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2016.