Maybe Look Closer? Detecting Trolling Prone Images on Instagram

Hitkul
IIIT-Delhi
Delhi, India
hitkuli@iiitd.ac.in

Rajiv Ratn Shah
IIIT-Delhi
Delhi, India
rajivratn@iiitd.ac.in

Ponnurangam Kumaraguru
IIIT-Delhi
Delhi, India
pk@iiitd.ac.in

Shin'ichi Satoh
National Institute of Informatics
Tokyo, Japan
satoh@nii.ac.jp

***Abstract*—Improvement in network infrastructure and smartphones have made images based social media platforms like Instagram and Flickr popular. The visual medium of communication has also led to an alarming increase in trolling incidents on social media. Though it is crucial to automatically detect trolling incidents on social media, in this paper, we look at the problem from the eye of prevention rather than detection. A system that can recognize trolling prone images can issue a warning to users before the content is posted online and prevent potential trolling incidents. We attempt to make a supervised classifier to detect trolling prone images and discuss why the conventional state-of-the-art image classification method does not work well for this task. We also provide an extensive analysis of trolling patterns in images from Instagram, discuss challenges and possible future paths in detail.**

***Keywords*-Social Media Computing; Computer Vision; Trolling; Instagram; Deep Learning**

I. Introduction

Social Media's astonishing growth [1] and its transformation into the new primary source of news consumption, marketing, advertising, earning [2] are clear indicators of how social media has become an essential part of our lives. However, this rapid growth of users social media has resulted in widespread propagation of damaging acts such as Trolling on the Internet. Trolling is defined as an act of "creating discord on the Internet by starting quarrels or upsetting people by posting inflammatory or off-topic messages in an online community." [3]. Trolling started as a means of self-amusement and entertainment of the troll, better known as *"for the lolz"* in social media language [4]. However, in recent times, trolling has grown from a tool of self-amusement to a tool for spreading propaganda [5].

It has been reported by the National Crime Prevention Council of the United States that about 40% of the population affected from trolling are teenagers [6]. Acts like trolling has resulted in mental illness, emotional trauma, psychological disorder and even suicidal intentions in individuals [7], [8].

Pater et al. [9] shows that the popular medium used by teenagers for communication includes apps such as Instagram, Vine and Snapchat. Such platforms are popular among teens due to their image and video sharing features. Advancements in camera qualities of portable digital cameras and smartphones have assisted in the rapid growth of images based on social media platforms. This resulted in an increase in visual content transfer by 70% on the web. This also increased trolling activities on such platforms to a vast extent. Works [10], [11], [12] indicates a noteworthy increase in using images and videos content for trolling, which led to much more prominent trolls [13]. That said, the majority of research in this domain concentrates on text modalities. There is a need to build systems which can leverage the information and signals presented by the visual modality also to tackle the problem of trolling on social media successfully.

Past research done in this area aims at detecting trolling after it has been posted on Instagram using available comment text, meta information and images for performing troll detection. Though such systems can successfully differentiate between trolls and non-trolls post with a certain degree of accuracy, it will not revert the damage caused. We want to take the stand of "Prevention is better than cure" in this study. As indicated by [14] there is a need to detect images which have a high probability of getting trolled at an early stage before it gets posted on social media. Such kind of system can warn users of potential future trolling events. For this task it is important to do the detection using only image features, because comment text is not available before posting of image.

Few studies have tried to leverage image modality for trolling detection in the past. They follow a supervised image classification approach using the favoured method of extracting features for classification using a convolutional neural network (CNN) models pretrained on ImageNet dataset. However, a close look into the results suggests that this method does not perform adequately for trolling prone image detection problem. The two main contribution of this paper are 1) highlight why current state-of-the-art image classification models fail to detect trolling prone images, 2) what kind of visual features are required to solve this problem. We do a extensive evaluation on why these models perform poorly and what is required to build systems which can predict the probability of an image being trolled.

The rest of the paper is organised as follows. Section II discusses the related work done in the domain in trolling detection with an emphasis on studies using the image data. This is followed by a discussion of the proposed methodol-

ogy in Section III. In Section IV, we present our experiment setup and detailed analysis of the observations. This is followed by Section V, that discusses the limitations and pointers for future work in this field. Finally, we conclude the paper with Section VI.

II. Related Work

Previous research in offensive content detection has revolved in the realm of Natural Language Processing with the user and metadata features as supporting features. Early research focuses on detecting offensive content on platforms like YouTube and Twitter using hand-crafted text features like token level n-gram, char level n-gram [15], [16], [17], [18], [19], [20], length of sentences [21], [18], capitalisation [22], [23], [18], [20] and sentiment [22], [17], [20]. Text features are many times supplemented with user information and metadata like age of the account [23], [24], [21], the number of followers/followee [25], [24], [22], [26], and profanity in username [22], [17]. More recent literature uses platforms like Reddit, Twitter and Instagram as a source. For classification task, text is represented using dense word vectors [22], [17], [18], [26], [27], [28] like Word2Vec [29]. Deep learning models like CNN [30] and CNN-GRU [27] are used for feature extraction and classification. Though a large amount of research has been done in detecting offensive content using NLP, performance comparison among the studies is not possible as the majority of studies are conducted on proprietary datasets.

One of the severe disadvantages of social media text data is the limitation of text length and format. Twitter is the most widely researched social media platform due to its friendly API and active user base, but it poses a total char restriction of 280 char (140 char in the past). Due to this char limitation and informal nature of the platform, people tend to write in short forms and abbreviation, which further hinders the information extraction from text. However, the upside of social media platform is the rich metadata and availability of multiple modalities. Information extracted from metadata and multiple modalities can be combined can lead to better results.

Multiple instances in literature have tried to use other modalities of information with text feature for detecting offensive content. Dinakar et al. [14] used knowledge graph for detecting subtle and sarcastic trolling attempts, Potha and Maragoudakis [31] modeled the problem as a time series prediction problem. Cheng et al. [32] formed multimodal graph representations using content, user and metadata information which is in turn used for classification. Though studies trying to detect trolling prone images using only visual features are non-existence. Work [25], [33] and [34] have used visual features as a supplement to text feature in the classification task.

Hosseinmardi et al. [34] created a dataset of 2000 (1500 normal and 500 trolling) Instagram session. A session includes an image, caption, associated comments, likes and user information. This dataset is publicly available on request. They analysed this dataset to find correlation between trolling and properties of session like linguistic features (word count, presence of first/second person noun and presence of swear word), temporal and graph properties (time of comment), Psychological features (content belong to social, family anger, sadness, sexual, etc.) and visual features (object present in image such as text, person, bike, food etc.).

Zhong et al. [25] created their own dataset with 3000 Instagram posts. However, this dataset is not publicly available. They used a combination of text features (Bag of words, word2vec and caption LDA) and visual features (vectors created by clustering the output of VGG16 pretrained on ImageNet). However, the best performance was shown by textual features. They also performed another set of experiments with a focus on prevention rather than detection just like us. A variety of visual features like SIFT, color histogram, GIST and feature vector extracted by trained VGG was used. All the classic visual features led to an accuracy between 52% to 57%, which is very close to the random guess accuracy of 50%. The best performance was shown by a model using the caption and VGG feature vectors at 68.55%. However, the model that only used caption gives an accuracy of 68.09%. This proves that visual feature did not provide any significant help. Solo performance of VGG features was 59.82%.They do not discuss the poor performance of of there solo visual models.

Singh et al. [33] used high-level visual features like adult score, the median age of people, raciness extracted from Microsoft Project Oxford API along with standard text features to do a supervised classification. They show a combination of two modalities achieves the best performance. However, the solo visual performance was still significantly lower than the solo text performance.

III. Methodology

Our study is divided into three phases. In the first phase, we do supervised image classification on features extracted from pretrained CNN . Subsequently, in the second and third phase, we evaluate the correlation of image class (trolling or normal) with low level and high-level object information of the image, respectively.

A. Supervised Learning Image Classification

Past literature for this problem has shown that features extracted using pretrained deep learning models perform better than classic computer vision features like color histogram, SIFT and GIST [25]. We used various CNN architecture pretrained on ImageNet dataset and a custom CNN trained from scratch for features extraction.

Extracted features are then passed into a fully connected layer for classification. We also tried to use SVM with RBF

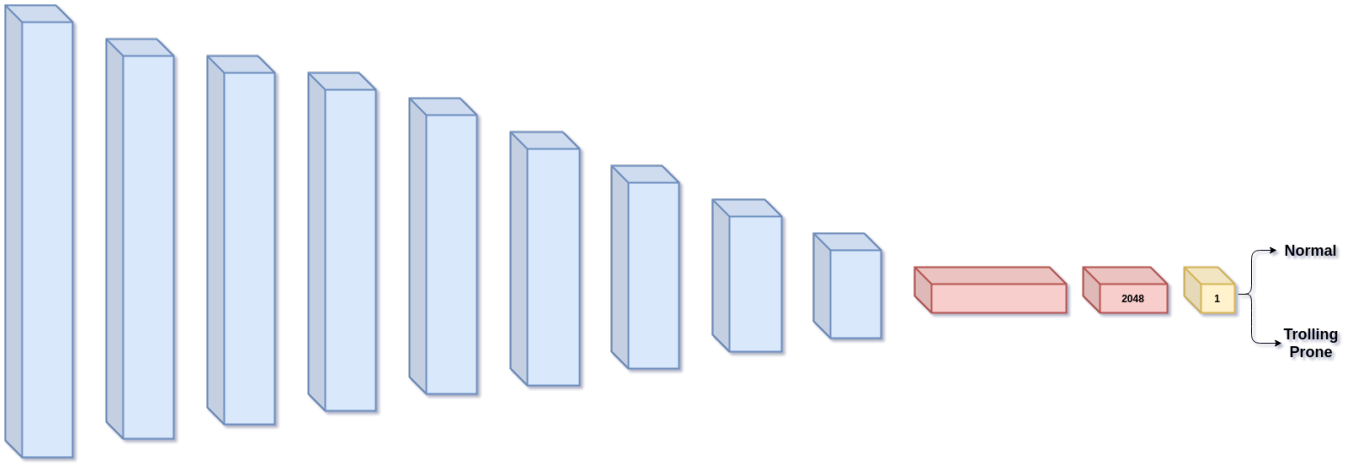


Figure 1: Supervised image classification architecture.

kernel as a classifier. A full outline of our classification model architecture is shown in Figure 1.

B. Low-Level Object Analysis

In the second phase, we compare low-level information about objects present in an image with the class of the image to spot any patterns in the trolling behaviour. By low-level object information, we mean information about the primary object names, e.g. person, text, clothes, drugs and cars. We do not capture any information about the physical attributes of the object.

We took any object which was present more than five times in the dataset.The objects which occur less than five times are clustered together in "other" category. This was done to reduce total number of unique objects, as samples occurring less that five time would not be able to give any assertive statistics. In total, we had 42 unique objects. For every object, we count the number of times it has been present in the image which was trolled/normal, and plot a stacked bar chart for the analysis. An outline of the process is shown in Figure 2

C. High-Level Object Analysis

This analysis is very similar to the one discussed in the previous section. However, In this phase, we labelled the images to capture the higher level of information about the object present. For instance, instead of just labelling text, we specify if the type text present, e.g. religious, motivational etc. We capture information about the physical attribute and properties of people and objects. Since the low-level object information was already present in the dataset, higher level attributes about the objects were added by manual labeling.

Labelling was done by one graduate and three undergraduate students. Each sample was annotated by two annotators. Inter agreement score was 91.75%, conflicts are resolved by discussion.

After labelling, for each attribute, we count the number of times it was present in a trolled image and a normal image and then plotted a stacked bar chart for the analysis. An outline of the process is shown in Figure 2

IV. Experiment and Observations

In this section, we would discuss the experiment setup and observations.

A. Dataset

We use the publicly available Instagram dataset introduced in the work [34]. Original dataset consisted of 1954 Instagram media sessions. A session includes the image, and associated metadata like comments, number of likes, caption and profile statistics of the uploader. A list of low-level objects is also available for each image. 566 sessions belong to the Trolling class, and 1388 belong to the Normal class.

At the time of experiments, only 1098 media sessions out of original 1954 were available. 557 of these are of class Trolled, and 541 belong to Normal class. Though the dataset contains useful metadata for each image, we are simulating a case where the classification of images is done before they are uploaded of Instagram. Hence, in all experiments, only images and available object level information for the images is used.

To further facilitate the analysis, high-level object information for each image was added to the dataset manually by a graduate student. A split of 80, 10, and 10, respectively, are used for train, validation and test sets in experiments.

B. Supervised Classification

CNN architectures pretrained on ImageNet dataset are used for feature extraction. We experiment with VGG16, VGG19 [35], ResNet50 [36], InceptionV3 [37] and InceptionResNetV2 [38] models. We also trained a CNN with five

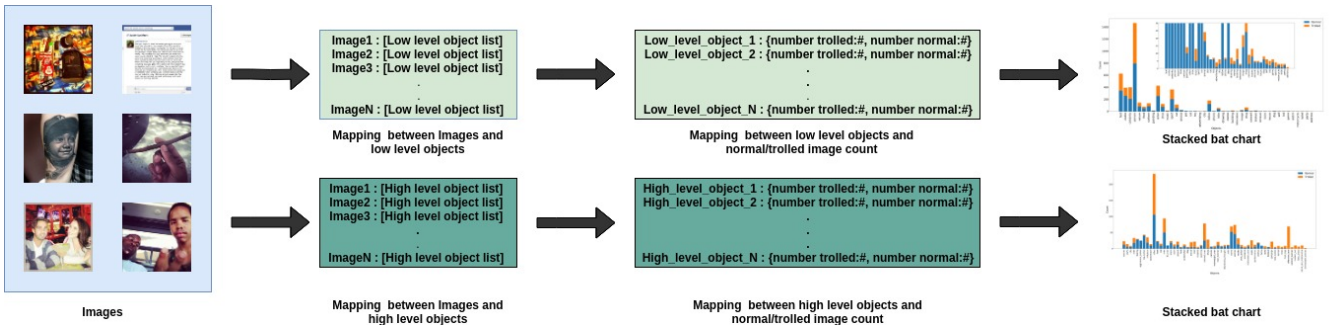


Figure 2: Object analysis architecture.

Table I: Results of supervised image classification.

Model	Validation Accuracy	Test Accuracy
VGG16	63.54	**61.81**
SVM (VGG16 features)	57.79	**61.81**
SVM (VGG19 features)	60.55	58.18
Inception Resnet V2	63.54	55.45
Inception V3	**65.62**	53.63
VGG19	63.5	51.81
Custom	54.16	51.81
ResNet 50	52.08	50.90
Zero Rule baseline	50.72	

blocks of convolution, max polling and batch normalization layers from scratch.

Extracted feature vectors are pass in a fully connected layer of size 2048 neurons, 50% dropout probability and relu [39] activation function. Classification layer has one neuron and sigmoid activation function. We also experimented with a RBF kernel SVM as classifier.

All the models are trained using data augmentation with a batch size of 32 and the learning rate of 10^{-5}. The number of hidden layers, number of neurons, dropout probability and batch size, are decided using the random search and learning rate is selected using a cyclical learning rate method as described in [40].

1) Results: To the best of our knowledge [33] is the only study which conducts visual feature based experiment on the same corpus. However, they only used 699 sessions from the dataset, and hence, a direct comparison is not feasible. Zhong et al. [25] performed similar experiments on there own Instagram dataset, which is not publicly available. Though a direct comparison is not possible, we can still analyse their results for patterns.

In Zhong et al. [25] experiments, visual features extracted from pretrained deep learning models outperform classic visual features like a colour histogram. However, there model using deep learning extracted features show extreme overfitting. Even after rigorous parameter fine-tuning, a similar pattern is observed in our experiments. Our training improves continuously as training time increases. However, validation performance remains steady on an average. A small validation dataset causes more volatility in results. The best test accuracy is 61.81%, which is just 10% better than baseline of zeroR classifier [41]. Absolute performance scores for all our models are given in Table I.

The critical question is to understand why the state-of-the-art image classification method perform not so good at this problem, as for many other computer vision task this method can deliver human level performance. Our, next set of experiments are performed in hope to analyse the shortcoming of our classification model.

C. Low-Level Object Analysis

This experiment tries to unroll why pretrained ImageNet feature do not perform well for detecting trolling prone images. Pretrained ImageNet models are trained to classify general objects like human, car and food. To check if such object patterns can be helpful, we conduct a simple experiment of checking frequency of an object towards each class. Our dataset already had these low-level object information for each image, similar to what ImageNet models are trained to identify. Figure 3 shows our analysis graph.

1) Observations: As it can be observed from the graph in Figure 3, most of the objects have a balanced ration of trolled and normal images. Few classes which show an inclination towards trolling are drugs, concerts, watches and advertisement. However, all these objects together only make 7% of our entire dataset. This proves that predicting trolling prone images using general object information (like captured by pretrained ImageNet models) is hard.

Trolling as a phenomenon is rarely the function of an object present in the image. It is more dependent on the characteristics of the objects. For instance, two of the most frequent objects in our dataset are person and text written in the image. However, trolling is not merely dependent on the presence of a person or text; it is a function of the appearance/action of this person or what is written in the text.

Proven that by nature, the problem of detecting trolling prone images is different from ImageNet classification task our best choice as recommended in [42] is to fine-tune the convolutional base on our dataset. Though, due to the tiny

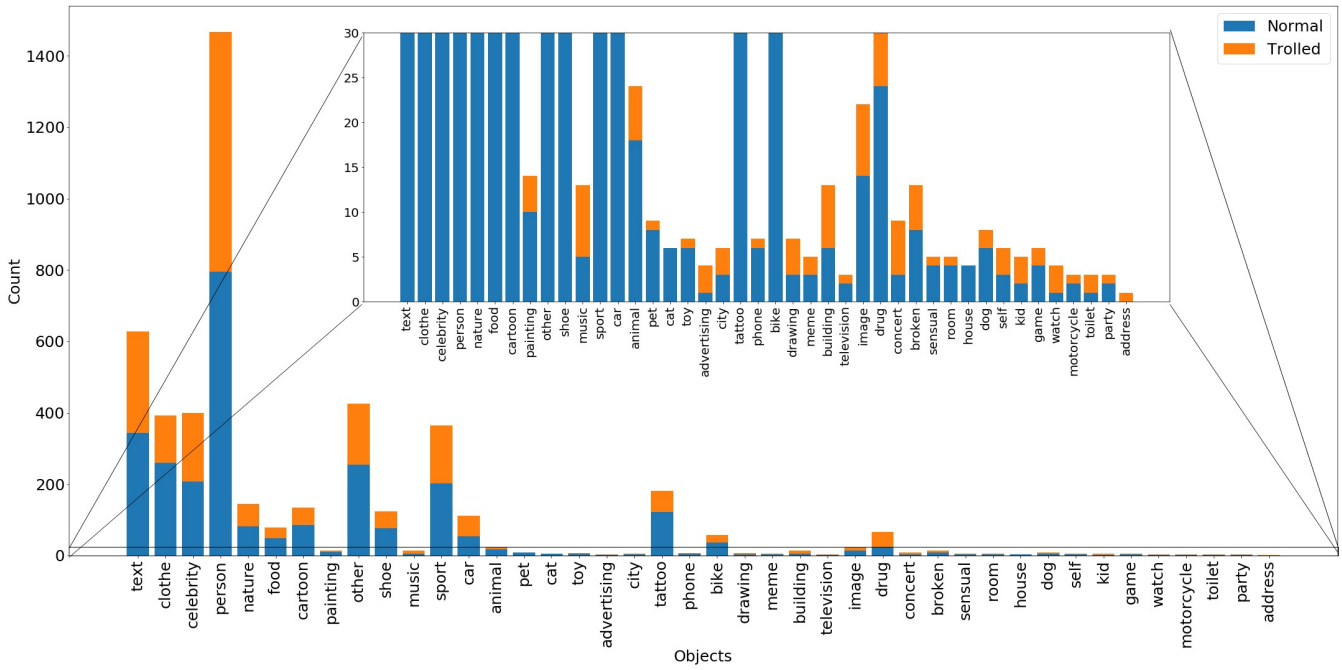


Figure 3: Frequency of Low-level object information towards each class.

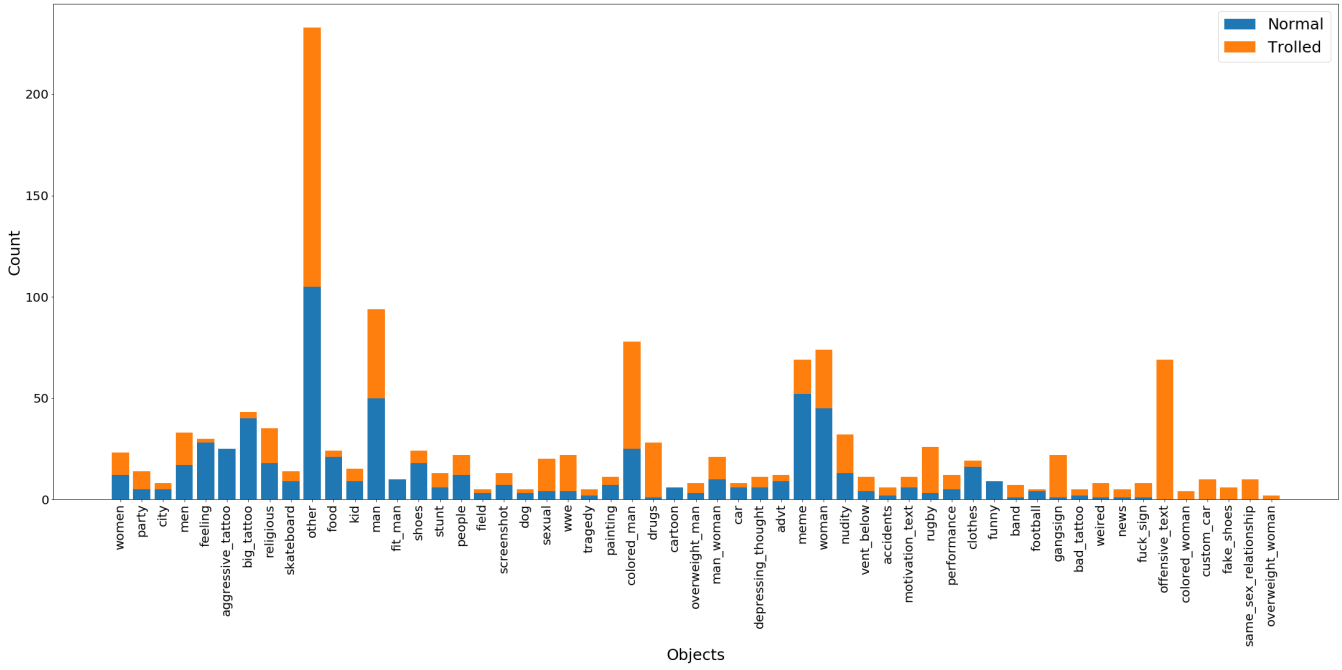


Figure 4: Frequency of High-level object information towards each class.

(a) Good quality big tattoo (Not Trolled).

(b) Bad quality big tattoo (Trolled).

Figure 5: Two Images of low-level object Tattoo. Looking at high level attributes in (b) the tattoo is of bad quality and got trolled, whereas (a) is a tattoo of good quality and did not got trolled.

Table II: Top 10 trolled high-level objects in our dataset.

High Level Object	Proportion Trolled
overweight_woman	1.0
same_sex_relationships	1.0
fake_shoes	1.0
custom_car	1.0
colored_woman	1.0
offensive_text	1.0
drugs	0.96
gang_signs	0.95
rugby	0.88
fuck_sign	0.87

Table III: Top 10 normal (not trolled) high-level objects in our dataset.

High Level Object	Proportion Normal
funny_text	1.0
cartoon	1.0
fit_man	1.0
aggressive_tattoo	1.0
fellings_text	0.93
big_tattoo	0.93
food	0.87
clothes	0.84
football	0.8
meme	0.75

size of our dataset that cannot be done. So for now we have to look for patterns in object properties. In the next phase of experiments, we would try to analysis that.

D. High-Level Object Analysis

In this experiment, we repeat the same analysis as explained in Section IV-C but with high-level object information. High-level object information contains properties and action of objects in the image. For instance, instead of a person, it would have labels like *man, woman* to indicate gender, *overweight_man/woman* to indicate obesity, *drugs_man/woman* to indicate consumption of drugs. Figure 4 shows the distribution of all the objects concerning classes trolling and normal.

1) Observations: Figure 4 shows some clear patterns of trolling. Some patterns which get trolled excessively are images with offensive text in them, bad quality tattoos, news, outlandish custom cars, unbranded shoes marked as branded (fake Nike shoes), rugby, drugs, WWE, sexual images, coloured or overweight women and images related to homosexual relationships. Table II and Table III shows our dataset‘s top 10 trolling and normal high-level object patterns respectively.

Vent below posts was trolled majority of times. It is a trend on social media when the user uploads an image with text “vent below” in it, which means he is asking the audience to write their feelings in the comments. On the contrast, some of the things which rarely get trolled are funny things like meme, clothes, fit people, good quality tattoos and cartoons. Figure 5, 6, 7 shows examples of how attribute difference of a object can be the deciding factor between getting trolled and not being trolled.

This analysis proves that it is possible to detect trolling prone images by looking at the properties and actions of the objects present in the image.

V. What is the catch? Limitations and Future Work

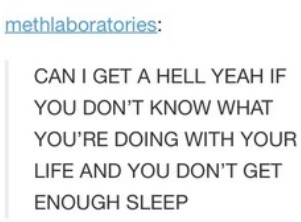
Results and analysis discussed in Section IV show that trolling prone images cannot be detected using ImageNet trained features, and we need to have higher level of information about images for this task. However, one of the limitations of this analysis is the sample size of the dataset.

Dataset used contains only 1098 images which is too small to represent a normalized sample set of Instagram

(a) A normal selfie (Not Trolled).

(b) Group of people showing gang signs (Trolled).

Figure 6: Two Images of low-level object Person. Looking at high level attributes in (b) people are showing gang signs, trying to look intimidating and got trolled, whereas (a) is a normal selfie and did not got trolled.



(a) General text expressing feelings (Not Trolled).

(b) Offensive language text (Trolled).

Figure 7: Two Images of low-level object Text. Looking at high level attributes in (b) text is written in a offensive language and got trolled, whereas (a) is a clean language text expressing feelings and did not got trolled.

content. Hence, the specific object properties extracted out of the analysis in Section 4 indicating high levels of trolling, may not be right always. For instance, our analysis shows that images related to to rugby are trolled approximately nine times out of ten, but this seems a bit unlikely. A better explanation is that the rugby-related account captured in the dataset we used has a high amount of trolling.

Though, this does not nullify the crux of this analysis that objects action and property information is the most valuable information when it comes to detecting trolling prone images for Instagram. Hence, any further research in this field should keep this analysis in mind and should try to eliminate our experiments shortcomings.

One of the primary requirement for solving this problem and getting a more normalised sample set is to get a larger dataset. However, there are two significant challenges in this direction are: 1) finding another source of data as recently Instagram public data collection API was shut down [43]. 2) Labelling a large amount of data. Though trolling is very prevalent on social media, it still only makes up a few per cent of the whole user-generated content on social media. Hence for acquiring a decent amount of labelled trolling data, a huge corpus will need to be labelled manually. Hopefully, with the improvements of natural language processing (NLP) models in future, labelling can be done automatically in future.

In terms of modelling, one of the first choices for future should be to fine-tune/retrain CNN models to extract relevant features for this task. However, this would need a dataset of magnitudes similar to ImageNet dataset. Another approach

can be to build models to detect each of the high level patterns mentioned and use a combination of them for final prediction.

An analysis of history of user and regular audience can also play a import role in will a image be trolled or not. For instance, images related to homosexuality is accepted in some part of the world where as it may get backlash in other parts of the world. Therefore, future work should consider using a combination of image features and meta features of audience.

VI. Conclusion

The expansion of social media, trolling has become one of the significant challenges in society today. It is easy availability to children makes it even more dangerous issues. With the growth of good quality portable cameras, image-based platforms like Instagram are the future of social media. Value of a system which can analyse and detect trolling prone images before they get posted on the internet is unprecedented as it can save people from immense metal stress. In this paper, we discuss how current state-of-the-art image classification methods are not well suited for this task, and we need a system which can understand the properties and actions of people and object in images. We lay out a detailed analysis of significant challenges in this field and some thoughts for future work.

Acknowledgment

Hitkul is partly supported by the NII International Internship program and MIDAS Lab.

Rajiv Ratn Shah is partly supported by the Infosys Center for AI, IIIT Delhi and ECRA Grant by SERB, Government of India.

References

[1] 10 social media statistics you need to know in 2019 [infographic]. [Online]. Available: https://www.oberlo.com/blog/social-media-marketing-statistics

[2] What is the real impact of social media? [Online]. Available: https://www.simplilearn.com/real-impact-social-media-article

[3] U. of Nebraska-Lincoln — Web Developer Network. Trolls and their impact on social media — james hanson. [Online]. Available: https://unlcms.unl.edu/engineering/james-hanson/trolls-and-their-impact-social-media

[4] C. Hardaker, "Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions," 2010.

[5] J. Aro, "The cyberspace war: propaganda and trolling as warfare tools," *European View*, vol. 15, no. 1, pp. 121–132, 2016.

[6] Stop bullying before it starts. [Online]. Available: http://www.ncpc.org/resources/files/pdf/bullying/cyberbullying.pdf

[7] L. Beckman, C. Hagquist, and L. Hellström, "Does the association with psychosomatic health problems differ between cyberbullying and traditional bullying?" *Emotional and behavioural difficulties*, vol. 17, no. 3-4, pp. 421–434, 2012.

[8] A. Sourander, A. B. Klomek, M. Ikonen, J. Lindroos, T. Luntamo, M. Koskelainen, T. Ristkari, and H. Helenius, "Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study," *Archives of general psychiatry*, vol. 67, no. 7, pp. 720–728, 2010.

[9] J. A. Pater, A. D. Miller, and E. D. Mynatt, "This digital life: A neighborhood-based study of adolescents' lives online," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 2305–2314.

[10] S. Hinduja and J. W. Patchin, *School climate 2.0: Preventing cyberbullying and sexting one classroom at a time*. Corwin Press, 2012.

[11] S. J. Seiler and J. N. Navarro, "Bullying on the pixel playground: Investigating risk factors of cyberbullying at the intersection of childrens online-offline social lives," *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, vol. 8, no. 4, 2014.

[12] S. Shariff, *Sexting and Cyberbullying*. Cambridge University Press, 2015.

[13] J. Kornblum, "Cyberbullying grows bigger and meaner with photos, video," *USA Today*, 2008.

[14] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, pp. 18:1–18:30, Sep. 2012. [Online]. Available: http://doi.acm.org/10.1145/2362394.2362400

[15] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[16] S. Malmasi and M. Zampieri, "Challenges in discriminating profanity from hate speech," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 30, no. 2, pp. 187–202, 2018.

[17] L. G. Mojica and V. Ng, "Modeling trolling in social media conversations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

[18] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 2016, pp. 145–153.

[19] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste, "Detection and fine-grained classification of cyberbullying events," in *Proceedings of the international conference recent advances in natural language processing*, 2015, pp. 672–680.

[20] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE Access*, vol. 6, pp. 13 825–13 835, 2018.

[21] M. Dadvar, D. Trieschnigg, and F. de Jong, "Experts and machines against bullies: A hybrid approach to detect cyberbullies," in *Canadian Conference on Artificial Intelligence*. Springer, 2014, pp. 275–281.

[22] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on twitter," in *Proceedings of the 2017 ACM on web science conference*. ACM, 2017, pp. 13–22.

[23] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *European Conference on Information Retrieval*. Springer, 2013, pp. 693–696.

[24] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network," *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016.

[25] H. Zhong, H. Li, A. C. Squicciarini, S. M. Rajtmajer, C. Griffin, D. J. Miller, and C. Caragea, "Content-driven detection of cyberbullying on the instagram social network." in *IJCAI*, 2016, pp. 3952–3958.

[26] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. Almeida, and W. Meira Jr, "Characterizing and detecting hateful users on twitter," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[27] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in *European Semantic Web Conference*. Springer, 2018, pp. 745–760.

[28] R. Kapoor, Y. Kumar, K. Rajput, R. R. Shah, P. Kumaraguru, and R. Zimmermann, "Mind your language: Abuse and offense detection for code-switched languages," *CoRR*, vol. abs/1809.08652, 2018. [Online]. Available: http://arxiv.org/abs/1809.08652

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[30] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *European Conference on Information Retrieval*. Springer, 2018, pp. 141–153.

[31] N. Potha and M. Maragoudakis, "Cyberbullying detection using time series modeling," in *2014 IEEE International Conference on Data Mining Workshop*. IEEE, 2014, pp. 373–382.

[32] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "Xbully: Cyberbullying detection within a multi-modal context," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 2019, pp. 339–347.

[33] V. K. Singh, S. Ghosh, and C. Jose, "Toward multimodal cyberbullying detection," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2017, pp. 2090–2099.

[34] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the instagram social network," in *International conference on social informatics*. Springer, 2015, pp. 49–66.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[38] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[39] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[40] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 464–472.

[41] N. Sajid, M. Afzal, M. A. Qadir, and S. Afgan Khan, "The insights of classification schemes," *Sindh University Research Journal*, vol. 45, pp. 145–150, 01 2013.

[42] Cs231n convolutional neural networks for visual recognition. [Online]. Available: http://cs231n.github.io/transfer-learning/

[43] Instagram developer documentation. [Online]. Available: https://www.instagram.com/developer/