Automated Detection of Substance Use-Related Social Media Posts Based on Image and Text Analysis

Arpita Roy, Anamika Paul, Hamed Pirsiavash, Shimei Pan
University of Maryland, Baltimore County
Baltimore, MD, 21250 USA
Email: {arpita2, rupa3, hpirsiav, shimei}@umbc.edu

***Abstract*—Nowadays, teens and young adults spend a significant amount of time on social media. According to the national survey of American attitudes on substance abuse, American teens who spend time on social media sites are at increased risk of smoking, drinking and illicit drug use. Reducing teens' exposure to substance use-related social media posts may help minimize their risk of future substance use and addiction. In this paper, we present a method for automated detection of substance use-related social media posts. With this technology, substance use-related content can be automatically filtered out from social media. To detect substance use related social media posts, we employ the state-of-the-art social media analytics that combines Neural Network-based image and text processing technologies. Our evaluation results demonstrate that image features derived using Convolutional Neural Network and textual features derived using neural document embedding are effective in identifying substance use-related social media posts.**

***Index Terms*—social media, substance use, illicit drug, teens, neural network, convolutional neural network, document embedding**

I. Introduction

With worldwide growth, social media has become increasingly more popular among people of all age groups. Although social media can help people to stay connected and updated on latest happening, its risks cannot be overlooked, especially for adolescents. This particular age group is among the most active users of social media. According to a large scale study by Common Sense Media in 2015, teens nowadays spend almost 2 hours per day on social network sites such as Instagram, Facebook and Snapchat. But a lack of adequate self-regulation and susceptibility to peer pressure make them vulnerable to behavior and mental problems such as depression, cyberbullying and substance abuse. For example, peer pressure plays an important role in teen's experimenting with substances. According to the US-based National Center on Addiction and Substance Abuse [1], 90% of teens are initially exposed to pictures of their peers drinking, using drugs or passing out on social media before they reach the age of 15 years old. It also found that 75% of teens ages 12 to 17 claim that seeing pictures on social media depicting their peers using drugs motivated them to mimic the behavior. As a result, teens who are exposed to these images are three times likelier to consume alcohol, and four times as likely to use marijuana. These teenagers are also more likely to have friends who abuse prescription and illegal drugs. It is also known that the younger someone is when he begins using an addictive substance, the more likely he is to become addicted. According to the same report, twenty-five percent of Americans who start using an addictive substance before 18 years of age become addicted, while only one out of every 25 Americans who starts using an addictive substance after the age of 21 becomes addicted. When teens become addicted, they develop health problems, fail in schools, lose motivation, and alienate families and friends. Since reducing early exposure to substance use-related content is one of the most effective ways to prevent addiction and substance abuse, technologies that automatically detect substance use related posts may play an important role in making social media a safer place for adolescents to explore and interact.

In this paper, we report our effort on automated detection of substance use related social media posts. We apply neural network-based image and text analytics to automatically examine the picture and text (including hashtags) in a social media post. Because of the recent opioid epidemic in the US, we focus on social media posts that depict illicit drugs. We use both the image and the text in a social media post because frequently image or text (including hashtags) alone is not sufficient to unambiguously determine whether a post is related to illicit drugs or not. For example, to avoid detection, frequently, illicit drug-related posts are tagged with benign popular hashtags like "Rihanna". In contrast, to attract attention, people tag normal life activities with illicit drug-related hashtags. By combining information from both image and text, we can maximize the system's ability in detecting illicit drug related social media posts.

The main contributions of this research include

1) Applying the state-of-the-art neural network-based image and text analysis technologies to identify substance use-related social media posts. So far, there has not been much work on automated detection of substance use-related social media posts. Our work represents an early step toward this direction.

2) Conducting comprehensive evaluations to demonstrate the effectiveness of the proposed method. Our best model achieved 90% prediction accuracy and 75% F-Measure, which are significantly better than models that use image or text features alone.

II. Related Work

So far, there is only limited research focusing on detecting substance use-related posts on social media. Among them, PREDOSE (PREscription Drug abuse Online Surveillance and Epidemiology) [2] was designed to facilitate the epidemiologic study of prescription drug abuse using social media. PREDOSE used web forum posts and domain knowledge modeled in a manually created Drug Abuse Ontology(DAO) to facilitate the extraction of semantic information from User Generated Content (UGC), through a combination of lexical, pattern-based and semantics-based techniques. In addition, [3] designed an automatic supervised classification technique to distinguish Twitter posts containing signals of medication abuse. They used different textual features for prediction such as word n-grams, abuse indicating terms, drug slangs and word clusters. [4] explored mentions of Adderall on Twitter to identify variations in volume around college exam periods. Most of these systems however used only text features except in [5] where robust face image analysis algorithms were used to extract drug users' demographics such as their age and gender.

In addition to substance use, there also exists work on detecting offensive people (cyber-bully) or content (e.g., profanity) on social media. Among them, [6] proposed a Lexical Syntactic Feature (LSF) architecture to identify potential offensive users in social media. They incorporated a user's writing style, structure and specific cyber bullying content as features to predict the user's likelihood to send offensive content. [7] proposed a semi-supervised approach for detecting profanity-related offensive content on Twitter. It exploited linguistic regularities in profane language via statistical topic modeling on a large Twitter corpus, and detected offensive tweets using these features. In [8], content-based and user-based features were used to improve the detection of cyber-bullying. Similarly, most of them used only textual features.

Finally, there is also a rich body of work on detecting offensive content in images. Although they are not specific to substance use or social media, they also employ image processing technologies. For example, many of the systems were designed to detect pornographic images. As a result, they tend to focus on detecting skin regions. Among them, [9] used the correlation between skin region and non-skin region to detect pornographic image. [10] first detected human-skin blobs. Special features were derived from these skin blobs (size, orientation, position, solidity, eccentricity, etc.) for image classification. Similarly, [11] extracted color and texture features from arbitrary-shaped segmented regions. Then Gaussian mixture models were built for skin and non-skin region classification. A skin map was produced based on the classification result. Eigenregion features were then used to describe the layout of skin regions in an image and pornographic images were detected according to the skin layout. Moreover, [12] used fuzzy classification to identify skin tones and employed shape recognition to match faces and other elements of the human body in detection of pornographic images. In [13], a two-stage detection method was used to take advantage of both content-based image retrieval and skin color analysis. In this work, first, content-based image retrieval was used to determine whether the image contains humans or not based on different color and shape features. Then a detailed skin color analysis was performed to determine whether the image is pornographic or benign. [14] adopted a visual bag-of-words (BoW) model to improve classification performance. The BoW model extracted the most common patches that exist on a set of training images. In addition, some of the recent work on detecting offensive content used neural network-based image processing techniques. For example, a Convolution Neural Network (CNN) was used to classify pornographic images in [15]. [16] used a fast and precise neural network model called Multilevel Sigmoidal Neural Network (MUSNN) for image analysis. It also exploited various color spaces for skin detection.

Most of the related work listed above used only a single type of features such as textual, image or user features. In contrast, we used a combination of textual and image features in our neural network classifier. To the best of our knowledge, a combination of neural network-based image and text analysis have not been used in detecting substance use on social media.

III. Dataset Description

Data Collection: Because of its popularity among young adults, in this research, we choose Instagram posts as our data source. Instagram has become one of the most visited social media platforms for sharing pictures, videos, and text. It has become the go-to platform for visual storytellers around the globe. As of December 2016, Instagram has 600+ million monthly active users. Instagram is most popular among teens and young adults. In the United States, more than half of Instagram users are between 18 and 29 years old. Globally, 41 percent of users are 24 years of age or younger. Instagram plays a major role in young adults' lives. It serves as a place where people know what their friends are doing and what is the latest trend. But there is a dark side of Instagram where people glamorize unhealthy lifestyles such as drug use and binge drinking. A disturbing community of drug abusers seemingly glorifying drug use has sprung up on Instagram. There are posts depicting drugs and people who unabashedly consume them. Beyond pictures of people doing drugs, many drug dealers are turning to Instagram to advertise their products. Even though Instagram blocked a handful of drug related hashtags and posts from appearing in search results, it is not sufficient. Because new drug street names are created frequently to avoid detection, drug related hashtags are constantly evolving, which makes it difficult to identify drug-related posts simply based on hashtags. In this research, we employ the state-of-the-art image

and text analysis technologies to accurately identify substance use-related posts on Instagram.

We used hashtag-based search to collect a set of Instagram posts. The seed hashtags include the official and street names of illicit drugs, which can be found on the website of the National Institute on Drug Abuse (NIDA) [1]. Then we used these hashtags to retrieve relevant Instagram posts. We expanded the list of hashtags by extracting new hashtags that are frequently occurred in the retrieved posts. We retrieved additional posts based on the new hashtags. We bootstrapped the post retrieval and hashtag augmentation process a few times until the dataset became stable.

Data Annotation: For data annotation, we created a web interface which allows multiple annotators to label the data simultaneously, The interface displays the image, text as well as hashtags associated with an Instagram post. Based on all the information associated with a post, an annotator needed to determine whether a post is related to illicit drugs or not. Each post is annotated by two annotators. We only keep those posts where both annotators agree on the labels. Figure 1 shows a few samples in our dataset. In total, there are 100,500 Instagram posts in our dataset. Among them, 98,000 posts have ground truth labels (when there is an agreement between the annotators). The distribution of the labels is skewed. Among those posts with ground truth labels, 20% are positive and 80% are negative. Table I shows some statistics of the dataset.

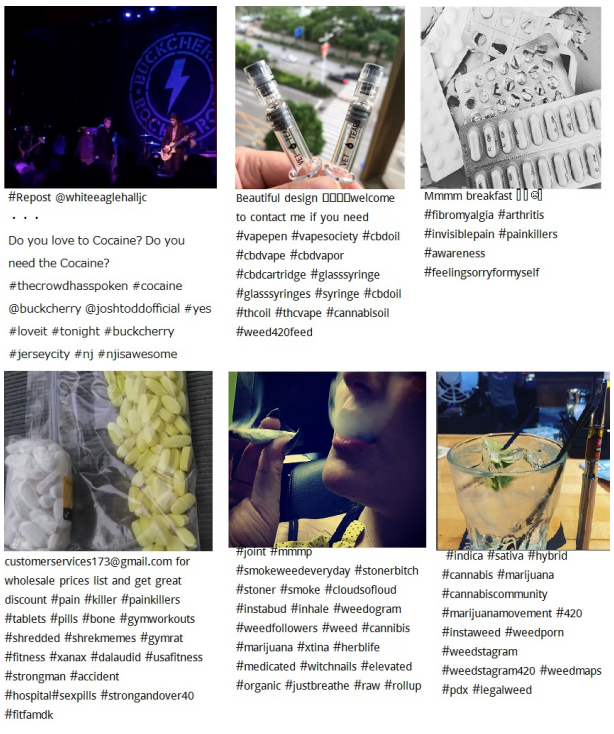

Fig. 1. A Sample of Collected Instagram Posts

[1]https://www.drugabuse.gov/drugs-abuse/commonly-abused-drugs-charts

Total posts	100,500
Posts with annotator agreement	98,000
Non Drug-related Posts	78,200
Drug Related Posts	19,800

TABLE I
DATASET SIZE AND DISTRIBUTION

IV. SYSTEM DESIGN

To detect social media posts related to illicit drugs, we need to first extract features from social media posts and then build a classifier that can classify those posts based on the extracted features. As Instagram often has both image and text (including hashtags) associated with most posts, we need to first extract both image features and text features. Then we combine these features in multimodal classification. In this research, we focus on neural network-based image and text feature extraction methods since most of the recent advancement in image and text processing employs such techniques. Figure 2 shows the system architecture. It contains three main parts (1) Image feature learning with Convolutional Neural Network (CNN) [17], (2) Textual Feature Learning with document embedding (Doc2Vec) [18] and (3) Neural Network Classifier. For image feature learning, we used CNN because it has been proven to be very effective in image feature extraction, image recognition and image classification ([19], [20], [21], [22]). CNN processes an image layer-by-layer in a hierarchical fashion, which is similar to how the human brains process visual input. Each layer of a CNN is trained to recognize image features with different levels of abstraction. For example, the first layer of a CNN typically recognizes only low-level image features such as edges and corners. The second layer combines the edges and corners to identify higher level features such as shapes. The next layers may extract more higher-level features such as faces. The last layer of CNN typically performs object classification based on annotated examples. The second part of the system performs textual feature extraction. Since each text post frequently contains multiple words and hashtags, we treat all the words and hashtags in a social media post as a document. Then we use document embedding trained using Doc2Vec to learn a feature representation for such a document. We choose Doc2Vec to extract text features because Doc2Vec can learn a dense vector representation for a document more effectively than traditional textual feature representations such as n-grams and bag-of-word models. Since Doc2Vec is capable of capturing semantic relationships between words and sentences, it was shown to be more effective in learning text features ([23], [24], [25], [26]). The third part of our system is a classifier that uses the extracted multimodal image and text features as its input and produces a class label (i.e., either drug-related or not) as its output. In the following, we provide more technical details about each part.

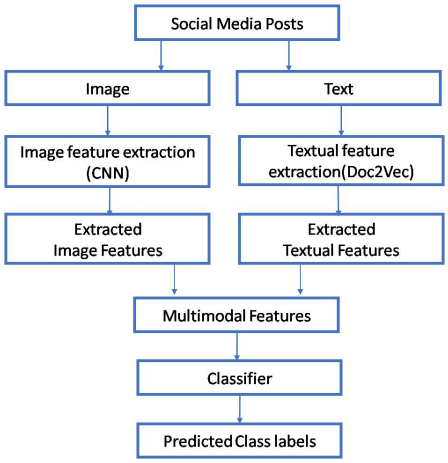

Fig. 2. System Architecture

Fully-Connected (FC) layer computes the class scores. As the name implies, each neuron in this layer is connected to all the neurons in the previous volume. The purpose of the Fully Connected layer is to use high-level features learned from the convolutional and pooling layers to classify the input image into various classes based on given training examples.

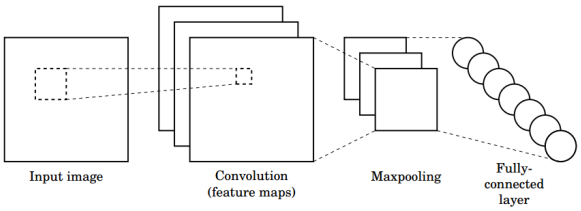

Fig. 3. Basic CNN Architecture

A. *Convolutional Neural Network (CNN)*

A CNN is a type of feed-forward artificial neural network comprised of one or more convolutional layers (often with a subsampling step), followed by one or more fully connected layers. Figure 3 shows the basic CNN architecture. CNN is very popular for image recognition. A CNN architecture is formed by a stack of distinct layers that transform the input features into classification scores through a differentiable function. A few distinct types of layers are commonly used in CNN.

INPUT layer holds the raw pixel values of an image with fixed width, height and three color channels R,G,B.

CONVOLUTION layer is the core building block of a CNN. The layer's parameters consist of a set of learnable filters (or kernels), which have a small receptive field but extend through the full depth of the input volume. During the forward pass, each filter is convolved across the width and height of the input volume, computing the dot product between the entries of the filter and the input and producing a 2-dimensional activation map of that filter. As a result, the network learns filters that activate when it detects some specific type of feature at some spatial position in the input. Stacking the activation maps for all filters along the depth dimension forms the full output volume of the convolution layer. Every entry in the output volume can thus be interpreted as an output of a neuron that looks at a small region in the input and shares parameters with neurons in the same activation map.

RELU layer applies an element-wise activation function. It increases the nonlinear properties of the decision function and the overall network without affecting the receptive fields of the convolutional layer.

POOLING layer down-samples the image data extracted by the convolutional layers to reduce the dimensionality of the feature map in order to decrease processing time. A commonly used pooling algorithm is max pooling, which extracts subregions of the feature map (e.g., 2x2-pixel tiles), keeps their maximum value, and discards all other values.

B. *document Embedding with Doc2Vec*

Doc2Vec is an unsupervised neural network-based machine learning algorithm that learns fixed-length feature representations from variable-length texts, such as sentences, paragraphs, and documents. Doc2Vec is an extension of one of the most popular word vector representation algorithms called Word2Vec [27]. Word2Vec employs a two-layer neural network that is trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. In the training process, context words are used to predict target word or vice verse. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to each another in the space. That means words with similar semantic meaning have similar vectors. These learned vectors explicitly encode many linguistic regularities and patterns. Many of these patterns and relations can be represented as linear translations in the vector space. For example, the result of a vector calculation vec(Madrid) - vec(Spain) + vec(France) is closer to vec(Paris) than to any other word vectors. This algorithm can overcome some of the main weaknesses associated with traditional text representations (e.g., bag-of-words models) such as missing semantic relations.

Since the text associated with a social media post typically contains multiple words or hashtags, Word2Vec is not appropriate since it only learns a representation for each word. Here, we need to learn a dense vector representation for all the words/hashtags in a post. Doc2Vec extends the original idea of word2vec and represents a sequence of words/hashtags as a dense vector which is optimized to predict any word/hashtag in the post. Details of two different frameworks of the Doc2Vec algorithm are given below.

Document Vector with Distributed Memory: In this framework, every document is mapped to a unique vector and every word is also mapped to a unique vector. The document vector and word vectors are averaged or concatenated to

predict the next word in a context. The contexts have a fixed-length and are sampled from a sliding window over the text. The document vector and word vectors are trained using stochastic gradient descent and the gradient is obtained via backpropagation. Figure 4 shows its architecture.

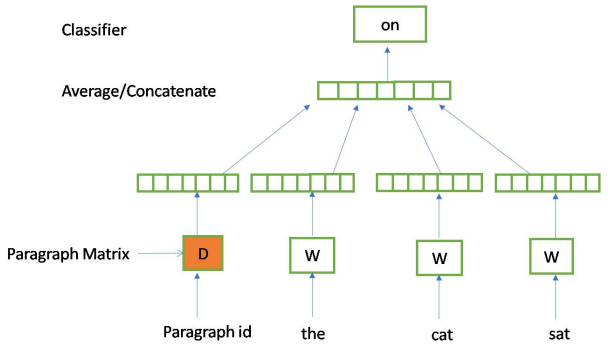

Fig. 4. Doc2Vec : Distributed Memory Model

Document Vector with Distributed Bag of Words: In this framework, word orders are ignored. Unlike the first Doc2Vec model, it ignores the context words in the input and forces the model to predict words randomly sampled from the input text. In reality, what this means is that at each iteration of stochastic gradient descent, we sample a text window. Then we sample a random word from the text window to predict using only the document Vector as its input features. Figure 5 shows its architecture.

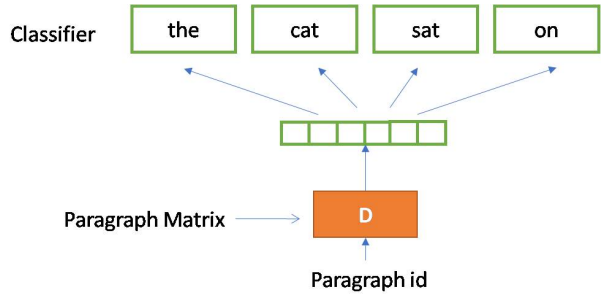

Fig. 5. Doc2Vec : Distributed Bag of Word Model

3. Neural Network Classifier : The third part of our system is a neural network classifier. Typically a multilayer perceptron [28] is used as a neural network classifier.

V. Experiments

In our experiments, we classified Instagram posts into two classes: (a) illicit drug related posts and (b) non-drug related posts. To get the best image and text features, we trained different models of CNN and Doc2Vec. To demonstrate the effectiveness of using combined features from both image and text, we also compared the models using multimodal features with those using only image or only text features. In all our experiments, we used 80% of posts as the training data and 20% as the test data.

A. Models with Image Features Only

In the following experiments, we varied the architecture and parameters of CNN to identify the model that is most effective in learning image features. The have evaluated a total of six image models.

A Simple CNN Trained from Scratch: First we tried a simple architecture of CNN. Our CNN consists of a simple stack of three convolutional layers with ReLU activation units, followed by a max-pooling layer. On top of it, we used two fully-connected layers. Our last layer includes a single unit with a sigmoid activation unit for binary classification. We trained the model by minimizing the binary cross entropy loss. We trained the model from scratch using only the annotated posts in our dataset. We have experimented with different numbers of fully connected layers and different numbers of neurons in each fully connected layer. We found that one hidden layer with 100 neurons worked best on our dataset. We call this model *Image Model 1*. Figure 6 shows the architecture of this model.

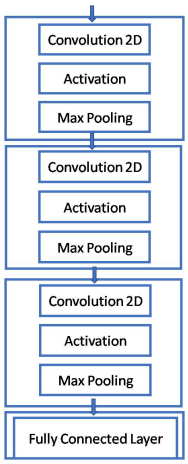

Fig. 6. A Simple CNN Architecture

Pre-Trained CNNs with Fixed Parameters: Since our dataset is relatively small, to improve system performance, next we tried a pre-trained CNN model to leverage the images in a large dataset. For this purpose, we used VGG16 [22], a pre-trained CNN model trained with 1.2 million images from ImageNet [29]. These images are categorized into 1000 classes, none of which is drug related. But the network may still capture general image features that can be useful in solving different image processing problems. In this experiment, we only instantiated the convolutional part of the model with the VGG16 parameters, which is everything up to the fully-connected layers. Running this model on our data gave us the "bottleneck features" from the VGG16 model. Then we trained a small fully-connected model on top of the stored "bottleneck features". During training, the "bottleneck features" worked as input and the weights in the fully connected layers were adjusted. We tried three configurations for the fully connected part: one has 1 hidden layer with 50 neurons (we call this *Image Model 2*); one has 1 hidden layer with 100 neurons (we

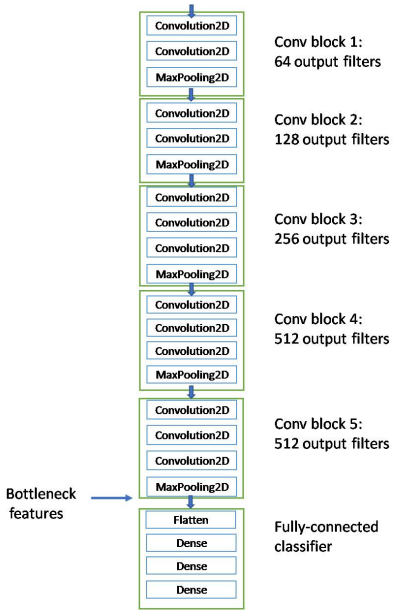

Fig. 7. VGG16 Architecture

call this *Image Model 3*) and one has two hidden layers with 250 and 50 neurons in each layer (we call this *Image Model 4*). Figure 7 shows the architecture of the VGG16 model.

Pre-Trained CNNs with Partial Parameter Tuning: Since illicit drug is not one of the categories used in the ImageNet dataset, it is possible that high level features captured in VGG16 are not optimized to detect drug-related images. In this experiment, we tried to fine tune the top layer of the pre-trained VGG16 models using our dataset. We started with the pre-trained VGG16 model. Then using our data, we fine-tuned the last convolutional block of the VGG16 model along with our classifier. The magnitude of the parameter updates stays small to avoid a significant change to the learned VGG 16 features. This model is called *Image Model 5*. Its fully connected part has 1 hidden layer with 100 neurons.

Pre-trained CNNs with Full Parameter Tuning: In this experiment, we initialize the CNN with VGG16 parameters trained from ImageNet data. Then we re-train the entire network of VGG16 using our data. This model is called *Image Model 6*. Its fully connected part has 1 hidden layer with 100 neurons.

B. Models with Text Features Only

We employed Doc2Vec to learn textual features to represent the words and hashtags social media posts. Each post is treated as a document by the Doc2Vec model. We experimented with both the distributed bag of word model and the distributed memory model. For each model, we also varied the feature dimensions and context window size. In total, we have tried five different configurations. *Text Model 1* is a distributed bag of word model with a vector dimension of 100 and a context window size of 10. *Text Model 2* is a distributed memory model with the same vector dimension and context window size. *Text Model 3* is a distributed memory model with 200 dimensions and a context window size of 10. *Text model 4* is a distributed memory model with 300 dimensions and a context window size of 10. *Text model 5* is also a distributed memory model with 300 dimensions and a context window size of 5.

C. Multimodal Classification

Finally, we combined both image and text features to train a classifier that can identify illicit drug-related social media posts. For this purpose, we combined the features from the best image and the best text models. This fully connected part of the Combined Model has 1 hidden layer with 100 neurons.

VI. Results

In this section, we present the results from different experiments. Our evaluation measures include accuracy, precision, recall and F1-measure.

A. Models Using Image Features Only

First we compare different models that classify Instagram posts based on image features only. Table II shows the results. In terms of accuracy, all the models achieved over 80% classification accuracy. The best model, *Image Model 6* that employs a pre-trained VGG 16 model plus full parameter tuning, achieved 87% classification accuracy. Since the class distribution in our dataset is skewed (e.g., about 80% negative cases), classification accuracy itself does not paint a complete picture. As shown in Table II, all the image models performed poorly based on model precision. The best model, *Image Model 6* only achieved only 46% precision. To measure the overall model performance based on the F1 measure, *Image Model 5* which employs pre-trained VGG16 plus partial parameter tuning performed the best with 46% F1 measure. We can draw two conclusions from these results (1) Pre-trained CNN models with parameter tuning are effective in boosting image-based classification performance. The two models that use pre-trained CNN models with either partial or full parameter tuning (*Image model 5* and *Image model 6*) performed the best in terms of accuracy and F1 measure. (2) Since the precision and F1 scores are pretty low, it seems image features alone are not adequate in identifying illicit drug-related posts.

B. Models Using Text Features Only

The performance of all the text-based models is summarized in Table III. Among all the text-based models, *Text Model 5* performed the best based on both accuracy and F1 measure. The model employs Doc2vec with Distributed Memory to learn a feature representation for all the text associated with a social media post. The dimension of the text vector is 300 and the context window size is set to 5. Similar to the models that only use image features, all the text models also have low precision and low F1 scores. The best text model, *Text Model 5* has only achieved a precision of 30% and a F1 of 40%. Thus, textual features alone are not adequate in identifying illicit drug-related social media posts.

C. Result with Both Image and Text features

Finally, we combined our best image and text features from the above in classification. This model with combined features achieved the highest accuracy of 90%, highest Precision of 74% and highest F1 measure of 75%. The improvement of precision and F1 measure is very significant. For example, compared with the best Image models, the combined model achieved a 60% increase of precision and 39% increase of F1 measure. Similarly, compared with the best text model, the combined model achieved a 147% increase of precision and 88% increase of F1 score. These results have demonstrated that image features and text features are very complementary to each other. When using text or image features alone, all the models performed poorly. When the text and image features are combined, the combined model outperformed the single models with a large margin.

Model	Accuracy	Precision	Recall	F1 measure
Image Model 1	83%	23%	65%	33%
Image Model 2	84%	29%	**79%**	42%
Image Model 3	85%	34%	76%	46%
Image Model 4	83%	20%	**79%**	31%
Image Model 5	86%	42%	76%	**54%**
Image Model 6	**87%**	**46%**	55%	50%

TABLE II
RESULTS OF IMAGE BASED MODELS

Model	Accuracy	Precision	Recall	F1 measure
Text Model 1	82%	18%	56%	27%
Text Model 2	**83%**	21%	58%	30%
Text Model 3	**83%**	23%	61%	33%
Text Model 4	**83%**	24%	**63%**	34%
Text Model 5	**83%**	**30%**	62%	**40%**

TABLE III
RESULTS OF TEXT BASED MODELS

Model	Accuracy	Precision	Recall	F1 measure
Combined Model	**90%**	**74%**	**77%**	**75%**

TABLE IV
RESULTS OF THE COMBINED MODEL

VII. VISUALIZING IMAGE CLASSIFICATION

We also want to gain some insight into the neural network models we built. Here we focus on CNN since its parameters are relatively easy to interpret. Here we use data visualization to examine which part of an image contributes to the classification. Class activation map is a simple technique to highlight the discriminative image regions used by a CNN. In other words, a class activation map (CAM) lets us see which regions in the image were relevant to the classification. Gradient-weighted Class Activation Mapping (Grad-CAM [30]) uses the gradients of a target concept, flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. Grad-CAM is applicable to a wide variety of CNN model-families . Figure 8 shows a few examples of the original images in our dataset and the generated class activation maps.

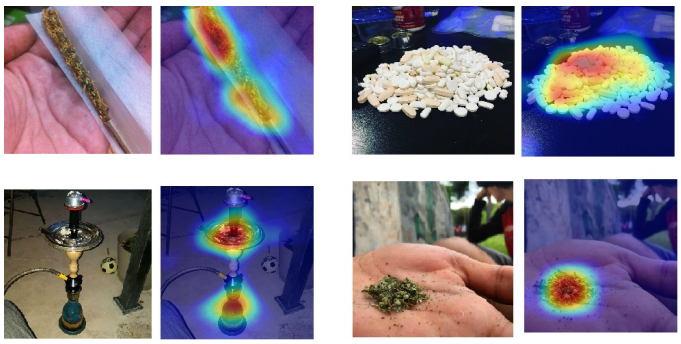

Fig. 8. Grad-Cam-based Visualization Applied to Samples in Our Dataset.

VIII. DISCUSSION AND FUTURE WORK

Even though our dataset has about 100K images, it is still not comparable to the size of the ImageNet dataset (with 1.2 million images). That is why the CNN models with pre-trained VGG16 works better than the one that uses only our dataset. We expect that with additional data, we can achieve even better performance.

The text in a social media post tends to be short, informal and noisy. For example, there is a lot of fluidity in how a given word is spelled on social media. It is also not uncommon that social media posts may contain a mixture of multiple languages and emojis. People also care less about grammar when writing on social media. All these make traditional NLP techniques such as syntactic parsing very difficult. In this research, we employ data-driven neural network based machine learning techniques such as Doc2Vec for text processing. The performance of these tools, however, depends on the quantity of the data. For example, it was difficult for Doc2Vec to produce good feature representations when the size of the dataset is relatively small. We expect that by collecting more data, we can improve the system performance even further. Currently, image and text features are combined together using a simple combining strategy (i.e. concatenation). In the future, we want to experiment with more advanced neural network-based multimodal feature learning strategies such as multimodal deep learning [31] and multi-view deep learning (Deep Canonical Correlation Analysis [32]).

IX. CONCLUSION

In this research, we employ the state-of-the-art image and text analytics to identify illicit drug-related social media posts. This technology can be used to filter out illicit drug-related posts from social media, which can significantly reduce the likelihood that young adults could be exposed to substance use related posts. We collected an Instagram dataset containing about 100K posts using a bootstrapping process for social media post retrieval and new hashtag identification. We used

both image features learned using Convolutional Neural Network and text features learned using Doc2Vec as the input to a neural network-based classifier. We describe a series of experiments to find the best model to automatically identify illicit drug-related social media posts. Our results indicate that combining information from both image and text is critical in accurately identifying illicit drug-related posts. Our best model achieved a 90% accuracy and a 75% F1-measure. This performance is significantly better than models that use only one type of features.

Illicit drug use and addiction are serious social problems. Image and text-based social media analytics plays an important role in finding a comprehensive solution to this problem. Our research represents an important step towards this direction.

Acknowledgment The authors would like to thank Arash Fallah, Yahya Almazni and Nirajkumar Lohbare for their contributions to the data annotation process.

REFERENCES

[1] T. Johnson, R. Shapiro, and R. Tourangeau, "National survey of american attitudes on substance abuse xvi: Teens and parents," *The National Center on Addiction and Substance Abuse*, vol. 2011, 2011.

[2] D. Cameron, G. A. Smith, R. Daniulaityte, A. P. Sheth, D. Dave, L. Chen, G. Anand, R. Carlson, K. Z. Watkins, and R. Falck, "Predose: a semantic web platform for drug abuse epidemiology using social media," *Journal of biomedical informatics*, vol. 46, no. 6, pp. 985–997, 2013.

[3] A. Sarker, K. OConnor, R. Ginn, M. Scotch, K. Smith, D. Malone, and G. Gonzalez, "Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from twitter," *Drug safety*, vol. 39, no. 3, p. 231, 2016.

[4] C. L. Hanson, S. H. Burton, C. Giraud-Carrier, J. H. West, M. D. Barnes, and B. Hansen, "Tweaking and tweeting: exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students," *Journal of medical Internet research*, vol. 15, no. 4, 2013.

[5] Y. Zhou, N. Sani, and J. Luo, "Fine-grained mining of illicit drug use patterns using social multimedia data from instagram," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1921–1930.

[6] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 2012, pp. 71–80.

[7] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 1980–1984.

[8] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context." in *ECIR*. Springer, 2013, pp. 693–696.

[9] Y.-C. Lin, H.-W. Tseng, and C.-S. Fuh, "Pornography detection using support vector machine," in *16th IPPR Conference on Computer Vision, Graphics and Image Processing (CVGIP 2003)*, vol. 19, 2003, pp. 123–130.

[10] J. Ruiz-del Solar, V. Castaneda, R. Verschae, R. Baeza-Yates, and F. Ortiz, "Characterizing objectionable image content (pornography and nude images) of specific web segments: Chile as a case study," in *Web Congress, 2005. LA-WEB 2005. Third Latin American*. IEEE, 2005, pp. 10–pp.

[11] Y. Xu, B. Li, X. Xue, and H. Lu, "Region-based pornographic image detection," in *Multimedia Signal Processing, 2005 IEEE 7th Workshop on*. IEEE, 2005, pp. 1–4.

[12] A. F. Drimbarean, P. M. Corcoran, M. Cuic, and V. Buzuloiu, "Image processing techniques to detect and filter objectionable images based on skin tone and shape recognition," in *Consumer Electronics, 2001. ICCE. International Conference on*. IEEE, 2001, pp. 278–279.

[13] B.-b. Liu, J.-y. Su, Z.-m. Lu, and Z. Li, "Pornographic images detection based on cbir and skin analysis," in *Semantics, Knowledge and Grid, 2008. SKG'08. Fourth International Conference on*. IEEE, 2008, pp. 487–488.

[14] T. Deselaers, L. Pimenidis, and H. Ney, "Bag-of-visual-words models for adult image classification and filtering," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[15] M. Moustafa, "Applying deep learning to classify pornographic images and videos," *arXiv preprint arXiv:1511.08899*, 2015.

[16] U. Sayed, S. Sadek, and B. Michaelis, "Two phases neural network-based system for pornographic image classification," in *Proceedings of 5th International Conference of Sciences of Electronic, Technologies of Information and Telecommunications (SETIT2009)*, 2009, pp. 1–6.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188–1196.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[21] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3642–3649.

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

[23] M. Campr and K. Ježek, "Comparing semantic models for evaluating automatic document summarization," in *International Conference on Text, Speech, and Dialogue*. Springer, 2015, pp. 252–260.

[24] J. H. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," *arXiv preprint arXiv:1607.05368*, 2016.

[25] Y. Belinkov, M. Mohtarami, S. Cyphers, and J. R. Glass, "Vectorslu: A continuous word vector approach to answer selection in community question answering systems." in *SemEval@ NAACL-HLT*, 2015, pp. 282–287.

[26] S. Lee, X. Jin, and W. Kim, "Sentiment classification for unlabeled dataset using doc2vec with jst," in *Proceedings of the 18th Annual International Conference on Electronic Commerce: e-Commerce in Smart connected World*. ACM, 2016, p. 28.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[28] S. Haykin and N. Network, "A comprehensive foundation," *Neural Networks*, vol. 2, no. 2004, p. 41, 2004.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[30] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *arXiv preprint arXiv:1610.02391*, 2016.

[31] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[32] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.