Detection of Illicit Drug Trafficking Events on Instagram: A Deep Multimodal Multilabel Learning Approach

Chuanbo Hu, Minglei Yin, Bin Liu, Xin Li∗
West Virginia University
Morgantown, WV, USA
cbh@wvu.edu.cn, my0033@mix.wvu.edu
bin.liu1@mail.wvu.edu,xin.li@mail.wvu.edu

Yanfang Ye
Case Western Reserve University
University of Notre Dame
Ohio/Indiana, USA
yanfang.ye@case.edu

ABSTRACT

Social media such as Instagram and Twitter have become important platforms for marketing and selling illicit drugs. Detection of online illicit drug trafficking has become critical to combat the online trade of illicit drugs. However, the legal status often varies spatially and temporally; even for the same drug, federal and state legislation can have different regulations about its legality. Meanwhile, more drug trafficking events are disguised as a novel form of advertising - commenting leading to information heterogeneity. Accordingly, accurate detection of illicit drug trafficking events (IDTEs) from social media has become even more challenging. In this work, we conduct the first systematic study on fine-grained detection of IDTEs on Instagram. We propose to take a deep multimodal multilabel learning (DMML) approach to detect IDTEs and demonstrate its effectiveness on a newly constructed dataset called multimodal IDTE (MM-IDTE). Specifically, our model takes text and image data as the input and combines multimodal information to predict multiple labels of illicit drugs. Inspired by the success of BERT, we have constructed a self-supervised multimodal bidirectional transformer by jointly fine-tuning pretrained text and image encoders. We have constructed a large-scale dataset MM-IDTE with manually annotated multiple drug labels to support fine-grained detection of illicit drugs. Extensive experimental results on the MM-IDTE dataset show that the proposed DMML methodology can accurately detect IDTEs even in the presence of special characters and style changes attempting to evade detection.

CCS CONCEPTS

• Information systems → Data mining; • Applied computing → Law, social and behavioral sciences.

KEYWORDS

drug trafficking; event detection; Instagram; multimodal bidirectional transformer; multilabel learning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

Chuanbo Hu, Minglei Yin, Bin Liu, Xin Li and Yanfang Ye. 2021. Detection of Illicit Drug Trafficking Events on Instagram: A Deep Multimodal Multilabel Learning Approach. In Proceedings of the 30th ACM Int’l Conf. on Information and Knowledge Management (CIKM ’21), November 1–5, 2021, Virtual Event, Australia. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3459637.3481908

1 INTRODUCTION

The co-evolution of cyberspace and human society has transformed the practice of illicit drug trade from the physical world to online platforms. Recent studies [14, 20, 27, 34, 35, 42, 51] have shown that major social media platforms, including Instagram, Twitter, and Facebook, have become a direct-to-consumer marketing tool for illegal drug dealers.

What makes the combat on online illicit drug trade even more challenging lies in the varying legal status of drugs. For instance, the use and possession of cannabis is illegal under federal law for any purpose in the US, but at the state level, policies regarding the medical and recreational use of cannabis vary greatly [5]. As of today, the recreational use of cannabis has been legalized in 15 states and decriminalized in another 16 states. Similarly, the legal status of club drugs varies according to the region and drug too. For example, some club drugs (e.g. cocaine) are almost always illegal; other club drugs (e.g. amphetamine or MDMA) are generally illegal unless with a lawful prescription from a doctor; other drugs (e.g. "poppers") are legal in some jurisdictions.

Detection of online illicit drug trafficking becomes a critical step to combat the illicit online trade of illicit drugs. However, it is challenging for the following reasons: (1) Inconsistency of drug legislation. Legal status of drugs varies according to the region and drug. Most recently, Oregon’s Measure 110 has decriminalized personal possession of small amounts of illegal drugs, such as cocaine, heroin, oxycodone, and meth; but the neighboring state of Washington has not. Such variation of legal status from state to state makes it difficult to draw a clear boundary between legal and illicit drug trade. (2) Information heterogeneity. The data sources related to drug trade involve both images and texts; meanwhile, the ways of advertising illicit drugs range from direct posting to indirect commenting (as a disguised form of advertising). More importantly, there is additional uncertainty arising from illicit drug dealer’s attempting to evade detection by different means (e.g., changing font styles, adding separators between letters, coining new street names of popular drugs). How to systematically combine these heterogeneous information has remained open. (3) Accuracy. When compared with normal users posting images and texts related to legal drug use, the portion of illicit drug trafficking is relatively
In this paper, we propose a deep multimodal multilabel learning (DMML) approach to detect the existence of multiple illicit drugs from suspect illicit drug trafficking events (IDTEs) on Instagram. As shown in Figure 1, a suspect IDTE is a user activity such as a post or a comment following a post on Instagram. Note that a suspect IDTE can be an initial post meant for marketing of illegal drugs. It can also be a comment following an initial post, in which drug trafficking information is added even though the original post does not contain any drug information. Unlike existing works on drug dealer detection [27, 42, 51] or drug use detection [14, 20, 34, 35] from aggregated information, we advocate to focus on detecting activities related to suspect IDTEs. This is because as the arms race between drug dealers and law enforcement evolves, more drug trafficking events are disguised as a novel form of advertising — commenting. Instead of directly posting some drug-related text or images (easily caught by the regulation), illicit drug dealers often indirectly advertise their products by adding a comment on the existing harmless post (refer to Fig. 1b). Note that such a piggyback strategy can be recursively applied, so the event of drug trafficking (drug-related comments) can be embedded at several levels under the original post. Our work is also different from existing works [14, 27, 42, 51] in the sense that our approach detects not only illicit drugs but also their specific types in each suspect IDTE. Such fine-grained detection of illicit drug trafficking becomes particularly important considering the inconsistency of drug legislation across different states.

Specifically, our model takes in text and image data associated with suspect IDTEs and composites the multimodal information to predict multiple labels of an illicit drug. Motivated by the latest advances in natural language processing - e.g., Bidirectional Encoder Representations from Transformers (BERT) [8], Vision-and-Language BERT (ViLBERT) [29], Learning Cross-Modality Encoder Representations from Transformers (LXMERT) [39]), we propose to develop a self-supervised multimodal bidirectional transformer (MMBT) by jointly fine-tuning pretrained text and image encoders. By projecting image embeddings to the text token space, we can employ self-attention over both modalities simultaneously, achieving more fine-grained multimodal fusion [22]. The combined multimodal feature is then passed to a multi-label learning module for predicting the type of multiple illicit drugs.

We have manually constructed a large-scale multimodal IDTE (MM-IDTE) dataset for the purpose of fine-grained illicit drug detection. Our MM-IDTE dataset, containing nearly 4,000 posts and more than 6,000 comments, represents the largest multimodal (text+image) illicit drug detection dataset so far. In particular, to construct such a large-scale dataset, we have designed an automatic data crawling system for Instagram that jointly uses hashtag and image information to guide the data collection. We have spent hundreds of hours on manually annotating each post (text and images) by multiple labels. Such multilabel ground truth has been verified by different people to ensure their consistency and accuracy. Extensive experimental results on the constructed dataset show that the proposed MMBT-based DMML approach can accurately detect IDTE. Both micro-/macro- precision and recall performance of our approach exceeds 0.90 on the test MM-IDTE dataset.

The key contributions of this paper are summarized as follows.

1. **We conduct the first systematic study on fine-grained detection of illicit drug trafficking events on Instagram. Unlike existing works on drug dealer or abuse detection, this work focuses on addressing the issue of commenting as a disguised form of advertising on Instagram.**

2. **We propose a deep multimodal multilabel learning (DMML) framework for detecting illicit drug trafficking events.** It is shown that the proposed MMBT-based approach can dramatically outperform unimodality and ad hoc multimodal fusion strategies. We have also experimentally compared different image encoders for MMBT-based fusion and found that MMBT based on ResNet50 and BERT achieves the best performance.

3. **We construct a large-scale MM-IDTE dataset for fine-grained illicit drug detection.** Toward this objective, we have developed an automatic hashtag-based data crawling system and a user-friendly data annotation system to support large-scale and multimodal data collection. The newly constructed MM-IDTE
dataset will be made publicly available to support the research related to illicit drug trafficking activities.

- We demonstrate the effectiveness of the proposed MMBT method on the MM-IDTE dataset. It is found that our method can successfully identify some challenging cases difficult for untrained eyes (e.g., special symbols and style changes attempting to evade detection). The developed system could facilitate the disruption of illicit drug trade by law enforcement.

2 RELATED WORK

2.1 Drug Abuse and Dealing Analysis
As far as we know, there has been limited work on tracking drug abuse and illicit drug trades from online data. Among these existing works, [4] analyzed the time and location patterns of drug use by mining Twitter data; network information of Instagram user timelines was used in [7] to monitor suspicious drug interaction activities; [52] and [42] analyzed Instagram data for tracking and identifying drug dealer accounts. More recently, machine learning and natural language processing techniques have been applied to combat prescription drug abuse [14, 20, 34, 35] and detect illicit drug dealers [27, 42, 51].

Our work is different from the previous works. First, our work focuses on suspected illicit drug trafficking events, while previous work focused on drug abuse or dealing with mining from aggregated information. Second, our method detects all potential illicit drugs in each suspected illicit drug trafficking event, while previous work either identified drug dealers or detected existing drug abuse. Third, we target at a fine-grained detection of different drugs as well as drug-related activities. Technically, we formulate our work as a multi-label learning problem [50], which is much more challenging than the binary classification in the previous works.

2.2 Multimodal Learning and Data Fusion
In many real-world problems, objects always involve multiple modalities. A modality refers to the way in which an object is represented. The goal of multimodal learning is to design a strategy to leverage the information from multiple modalities so that different sources of information can complement and enhance each other for a specific goal [2]. It usually involves a joint representations of different modalities and a way to fuse the representations to a composite multi-modal feature for the sake of the task in investigation. Multi-modal learning has enabled a wide range of applications such as multimedia content indexing and retrieval [6], image captioning [46], and visual question answering (VQA) [9].

Rapid advances of machine learning in recent years have also expedited the research in multimodal data fusion [47]. A straight way is to concatenate features from different modalities [23]. Bilinear pooling [9] based method was proposed to better capture the interactions between features in different modalities. A gated multimodal fusion module was proposed in [1] to find an intermediate representation based on a combination of data from different modalities. Its follow-up work has shown that fusion with discretized features outperforms text-only classification [23]. More recently, inspired by the success of Bidirectional Encoder Representations from Transformers (BERT) [8]), transformer-based multimodal data fusion has attracted increasingly more attention - e.g., multimodal bitransformer (MMBT) [22], Vision-and-Language BERT (ViLBERT) [29], Learning Cross-Modality Encoder Representations from Transformers (LXMERT) [39].

2.3 Multi-label Learning
Multi-label learning [50] targets at representing an object by a single instance but each object can be associated with a set of labels. In contrast to traditional supervised learning, the task of multilabel learning is to learn a function that can predict the proper label sets for unseen instances. Typically, the multilabel learning problem is transferred into other well-established learning settings such as binary classification, one-vs-all classification, or multiclass classification through the introduction of label powerset [40]. There are some efforts to adopt learning techniques such as low-dimensional label embedding methods [3], joint global and local approach [54], and joint learning of label-specific features and label correlations [48], to exploiting label correlations for multi-label learning.

Deep multi-label learning has also been recently studied for image classification [38] and in the special situation of extreme multi-label learning (XML) [28, 45].

3 ILLICIT DRUG TRAFFICKING EVENT DETECTION
In this section, we first formulate the problem of illicit drug trafficking event detection (IDTE), and then introduce the proposed deep multimodal multilabel learning (DMMML) approach.

3.1 Problem Formulation
Definition 3.1 (Illicit drug trafficking event). An illicit drug trafficking event (IDTE) is an event on Instagram that contains the marketing and selling of one or more defined illicit drugs.

In this paper, we consider the following nine common illicit drugs traded on Instagram: Marijuana, Codeine, Mdma, Xanax, Painkillers, Mushrooms, LSD, Cocaine, and other drugs.

Definition 3.2 (Suspect IDTE). A suspect IDTE is a user activity such as a post or a comment to a post on Instagram. It usually contains image and text information.

Figure 1 shows some examples of suspect IDTEs on Instagram. Note that a suspect IDTE can be an initial post (e.g., the post initialized by User A in Fig. 1a). It can also be comments following a post, in which drug trafficking information is added (e.g., comments by Users B and C in Fig. 1a). As shown in Fig. 1b, even the initial post does not include any drug-related information, illicit drug dealers can still advertise their products by adding a comment to an existing innocent post.

Given the above definitions, we can formally define the problem of illicit drug trafficking event detection as follows: The goal is to build an effective approach to detect the existence of illicit drugs such as cocaine and cannabis within each suspect IDTE i. Assume there are a total of C predefined illicit drugs under consideration. Let $y_{ic} = 1$, $c \in \{1, \ldots, C\}$ denotes suspect IDTE i contains drug $c$ and $y_{ic} = 0$ otherwise. Note that, it is also possible that the suspect IDTE i does not use any of the C illicit drugs. We add one more label $y_{i0} = 1$ to indicate the case when the suspect IDTE is drug-free. Then each suspect IDTE i can be represented as a vector $y_i = [y_{i0}, y_{i1}, y_{i2}, \ldots, y_{iC}] \in \{0, 1\}^{C+1}$. Each
suspect IDTE $i$ is associated with an image $x_i$, and a text comment $t_i$, which has $T$ tokens. Given a set of $N$ training instances $D = \{(x_1, t_1, y_1), (x_2, t_2, y_2), \ldots, (x_N, t_N, y_N)\}$, we aim to build a predictive model $f : [x, t] \rightarrow y$ to detect the existence and types of illicit drugs from new suspect IDTEs.

### 3.2 System Overview

Figure 2 shows the overview of our deep multimodal multilabel learning (DMML) approach to illicit drug trafficking activity detection. The model takes in text and image data associated with suspect IDTEs, and combines multimodal information to predict multiple labels of illicit drugs.

It is worth mentioning several salient features of our DMML approach before elaborating on its details. First, the complementary role played by text and image information has been recognized by previous work (e.g., [42]). However, it is unclear which modality contributes more to the detection especially when the problem of detection reaches fine-grained. Contrary to the findings reported in [42], we have found texts are more reliable than images for fine-grained drug classification. Second, we target at extracting multilabel information from IDTE to more accurately track the spatio-temporal dynamical patterns of different drugs. Such feature is important to address the issue of inconsistency of drug legislation across different states.

### 3.3 Multimodel Fusion via Bidirectional Transformer

Each suspect IDTE is associated with an image $x$, and a text comment $t$, which has $T$ tokens. Let $t = (w_1, \ldots, w_T)$. The presence of multiple information sources holds the promise to learn better feature representations for final detection of illicit drugs. Typically, text data $t$ is processed by a sequence model such as LSTM to form a text feature $\phi_t$, and image $x$ is processed using a pre-trained CNN model to form an image feature $\phi_x$. Then a fusion model is applied to combine the text and image features $\phi_{xt} = \text{fusion}(\phi_t, \phi_x)$. There are a wide range of fusion methods proposed [47] such as concatenation [23] and bilinear pooling [11, 18].

In this paper, we propose to adapt a bidirectional transformers approach to multimodal fusion. First, we found textual are more reliable than images for fine-grained drug classification. Second, as discussed in [41], multimodal models that composite features at late stages are prone to overfitting. The bidirectional transformers approach projects image embedding to the text token space to form better feature representations for final detection of illicit drugs.

#### 3.3.1 Transformers in Text and Vision

The idea of self-supervised embedding or transfer learning from pre-trained representations has been extensively explored in the literature of natural language processing (e.g., [25, 31]) and computer vision (e.g., [32, 36]). Most recently, the idea of fine-tuning self-supervised or semisupervised learning has revolutionized the field of natural language processing leading to breakthroughs such as BERT [8] and its variations (e.g., xlnet [43] and albert [26]). By applying the bidirectional training of transformer, a popular attention model, to language modelling, BERT learns contextual relations between words in a text more effectively.

The success of transformer architectures has rapidly leveraged to the field of computer vision [13]. By integrating self-attention with self-supervision, transformers can exploit long-range dependencies in the input domain, which make transformer-based representation more expressive. Since there is a minimal assumption about prior knowledge, pretrained transformers are particularly suitable for large-scale and unlabelled datasets. Thanks to the generalization of encoded features, learned representations can be fine-tuned by labelled data leading to excellent performance on various vision tasks [21].

#### 3.3.2 Multimodal Transformer

Inspired by the latest advances in multimodal transformers (e.g., ViLBERT [29], LXMERT [39]), we propose to develop a self-supervised multimodal bi-transformer (mimicking bidirectional transformers) jointly fine-tuning pretrained text and image encoders [22]. The basic idea is to use self-attention over both texts and images simultaneously, providing early and fine-grained multimodal fusion. It has been shown in previous work [22] that such conceptually simple strategy can work as effectively as more sophisticated multimodally pretrained ViLBERT models.

More specifically, we have constructed a multimodal bitransformer (MMBT) model (refer to Fig. 2) combining text-based self-supervised representations with image-based CNN architectures (e.g., ResNet [15]). By projecting image embedding to the text token space, we can employ self-attention over both modalities simultaneously, achieving more fine-grained multimodal fusion [22]. It should be noted that a salient feature of MMBT is the flexibility of plugging in different images and text encoders. For example, the ResNet-152 image encoder adopted in [22] can be readily replaced by a smaller ResNet-50 counterpart.

### 3.4 Multilabel Learning

The output of our model is the predicted probability vector $\hat{y}_i = \{\hat{y}_{i0}, \hat{y}_{i1}, \ldots, \hat{y}_{iC}\} \in \{0, 1\}^{C+1}$ for suspect IDTE $i$. This is a typical
example of multilabel learning, where each example represented by
a single instance is simultaneously associated with multiple class
labels [49]. Under the context of IDTE detection, multilabel learning
achieves fine-grained classification of multimodal data. To the best
of our knowledge, such fine-granularity classification has not been
considered in previous works of post-based drug dealer detection
[42]. Though there are different training strategies in multilabel
learning [54], we have adopted the binary cross-entropy (BCE) loss,
which has been widely applied in deep learning based multilabel
learning problems [28, 45]. The BCE loss is defined as
\[ \ell(\Theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=0}^{C} y_{ic} \log \hat{y}_{ic} + (1 - y_{ic}) \log(1 - \hat{y}_{ic}) \].

4 MULTIMODAL IDTE (MM-IDTE) DATASET CONSTRUCTION

In this section, we present a new multimodel dataset constructed
from Instagram; toward this objective, we will discuss our effort on
data crawling and data annotation, respectively.

Table 1: Comparison of our MM-IDTE dataset with several existing datasets for drug dealer and abuse detection.

| Study | Source     | Granularity(#) | classification       | Application         |
|-------|------------|----------------|----------------------|---------------------|
| [53]  | Instagram  | Post-level:2,362 | Binary               | Drug use pattern analysis |
| [42]  | Instagram  | Post-level:4,819  | Binary               | Drug dealer detection  |
| [30]  | Twitter    | Tweets-level:213,041 | Binary             | Drug dealer detection  |
| [16]  | Twitter    | Tweets-level:1,794  | Binary               | Drug abuse risk detection |
| [14]  | Instagram  | Post-level:369,000 | Binary               | Drug abuse risk detection |
| [27]  | Instagram  | Post-level:1,228   | Binary               | Drug dealer detection  |
| Ours  | Instagram  | IDTE-level: 4,648   | Multi-label         | IDTE detection       |

To achieve this objective, we have made a great effort on data
collection and construction of multimodel IDTE dataset in this project.
An integrated data collection and calibration platform has been
designed for automatic data collection/crawling and synchronized
multilabel data calibration, as shown in Figure 3. Three main mod-
ules in the platform have been implemented to realize suspect IDTE
data collection, data storage, and data calibration, respectively.

4.1 Data collection

The rationale underlying our data crawling scheme is still based on
hashtag-based search [12]. Hashtags on Instagram can help users
extend their reach, engage their audience, which can be attached to
posts, and become clickable phrases and topics with the # placed in
front of them. However, unlike [42] working with a fixed collection
of hashtags, we propose a data crawling algorithm that iteratively
expands the pool of hashtags for scaling up our search. Such expan-
sion of hashtags is guided by an intelligent pretrained AI model
(VGG-16 [37]) designed for drug image classification. By treating
drug-related hashtags and images as a pair of peer hidden variables,
our iterative crawling system aims at refining and updating the
collected multi-modal data in an Expectation-Maximization (EM)-like
manner. The detailed description of our data collection system
consists of the following four components.

1. Drug-related hashtags collection: A total of 200 drug-related
hashtags have been manually collected by domain experts
using the hashtag search API [10]. These hashtags contain
10 types of drugs (i.e., non-drug, marijuana, codeine, 3,4-
methylendioxy-methamphetamine (MDMA), xanax, painkiller,
psilocybin mushroom (hereinafter called mushroom), Lyser-
gic acid diethylamide (LSD), cocaine, other drugs), which are
widely trafficked on Instagram\(^1\). We have used this set of
hashtags as the initial starting point of our data collection.

2. Drug-related post detection. We search each post (which
includes an image and comments) with each drug-related
hashtag as input. A VGG-16 based binary classification model
[37] is pretrained to detect drug-related posts from the
accompanying image information. The image-based dataset for
model pretraining contains various types of drug-related im-
ages which are sources from Bing image search API (similar
to Google image search API adopted in [42]). If an image of a
post is detected by the model as being drug-related (positive),
we save its link for further processing.

3. Drug-related data collection. The detected posts were con-
verted and formalized into a universal json object to facilitate
the storage and retrieval. As post comments are sources from
several user accounts, we saved each post-related informa-
tion (including posted images and comments). Totally, 10,000
potential posts and 23,034 user homepage information were
collected as the initial dataset.

4. Drug-related hashtag update. New hashtags from each de-
tected post can be added into the list of drug-related hashtags.
We have also recorded the frequency of each hashtag in order
to track the most frequent ones. The system uses the new
hashtag (which have the highest frequent counts) in the next
iteration until the amount of collected data reaches a pre-
specified threshold (in this study, we have set the threshold
to be 1000 drug-dealer accounts).

4.2 Data annotation

Suspect IDTEs often contain various types of drugs, so we have de-
signed a multi-label annotation module, which contains Instagram
post data visualization and three information (i.e., Hashtag label,
image label and comment label) annotation. The three information
of any suspect IDTEs can be labeled by domain expert users through
10 optional categories. The labeled hashtags will update the drug-
related hashtag list weight to improve data collection efficiency.
Totally, 4,648 suspect IDTEs were labeled during the construction of
the experimental dataset, which contains 1,022 drug trafficking
posts and 1,406 unique drug dealer user accounts. The proportion
of each category label in the dataset is shown in Figure 4. Several
existing datasets have been designed for illicit drug dealing tracking,
as shown in Table 1. Compared with these existing datasets,
our dataset contains more user accounts and more diverse labels.

To ensure the consistency and accuracy of multiple labels, we
have asked different people to cross-validate the annotation results.
Several particular challenges we have identified through the manual
data annotation process include: 1) certain drugs (e.g., power-type)
have similar visual appearance and can be easily confused with

\(^1\)https://drugabuse.com/featured/instagram-drug-dealers/
which requires hundreds of hours of manual labor for cross-checking and validation.

Training and testing. We randomly split the dataset into a train-set (75%) and a testing set (25%). We have trained the models using the popular Adam optimization algorithm [33].

The following parameters are adopted in our setting: learning rate $\alpha = 2e^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. We opt to terminate the training after 50 epochs. All experiments are conducted using PyTorch on a workstation with one RTX 2080 GPU.

5 EXPERIMENTAL RESULTS

5.1 Experimental Setup

Data. The MM-IDTE dataset used in our experiment contains 4,648 manually labeled IDETE records. There are 10 labels including 9 drug labels (i.e., Marijuana, Codeine, Mdma, Xanax, Painkillers, Mushrooms, LSD, Cocaine, and other drugs) and a non-drug label when the IDTE is drug-free. Figure 4 shows the distributions of the 10 labels across the whole IDETE records.

Training and testing. We randomly split the dataset into a training set (75%) and a testing set (25%). We have trained the models using the popular Adam optimization algorithm [33].

The following parameters are adopted in our setting: learning rate $\alpha = 2e^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. We opt to terminate the training after 50 epochs. All experiments are conducted using PyTorch on a workstation with one RTX 2080 GPU.

5.2 Evaluation metrics

We evaluate all methods in terms of example-based and label-based multilabel classification measures.

Example-based metrics are defined by comparing the ground-truth label set $y = \{y_1, y_2, \ldots, y_C\}$ to the predicted label set $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_C\}$ on each test example, and then calculating the mean value across all test datasets. Subset accuracy is a strict metric that measures the fraction of correctly classified examples and requires an exact match of the predicted label set and the ground-truth label set. Subset accuracy is defined as subset accu. $=$ $\frac{1}{N} \sum_{i=1}^{N} \sum_{c=0}^{C} \mathbb{1}[y_i = \hat{y}_c]$, where $\mathbb{1}[y = \hat{y}]$ is an indicator function with value of 1 when $y = \hat{y}$, and value of 0 otherwise. Hamming loss evaluates how many labels are incorrectly predicted on average, and is defined as hamming loss $=$ $\frac{1}{N} \sum_{i=1}^{N} \frac{1}{C} \sum_{c=0}^{C} \mathbb{1}[y_i \neq \hat{y}_c]$.

Label-based metrics are defined by evaluating the prediction performance of each label separately, and then returning macro-averaged or micro-averaged metric values across all labels. Precision, recall, and $F_1$-measure are commonly used metrics. $F_1$-measure combines precision and recall, and is the harmonic mean of precision and recall. Specifically, for each $c$-th label in $\{0, 1, 2, \ldots, C\}$, we denote by $TP_c$, $FP_c$, $TN_c$, $FN_c$ the number of true positives, false positives, true negatives, and false negatives, respectively. Micro-averaged precision, recall, and $F_1$-measure are defined as follows: $P_{micro} = \frac{\sum_{c=0}^{C} TP_c}{\sum_{c=0}^{C} TP_c + FP_c}$, $R_{micro} = \frac{\sum_{c=0}^{C} TP_c}{\sum_{c=0}^{C} TP_c + FN_c}$, $F_{1micro} = \frac{\sum_{c=0}^{C} 2TP_c}{\sum_{c=0}^{C} 2TP_c + FP_c + FN_c}$. Macro-averaged precision, recall, and $F_1$-measure are defined as follows: $P_{macro} = \frac{1}{C} \sum_{c=0}^{C} \frac{TP_c}{TP_c + FP_c}$, $R_{macro} = \frac{1}{C} \sum_{c=0}^{C} \frac{TP_c}{TP_c + FN_c}$, $F_{1macro} = \frac{1}{C} \sum_{c=0}^{C} \frac{2TP_c}{2TP_c + FP_c + FN_c}$. Macro-averaging treats all labels equally while micro-averaging favors more frequent labels. High macro-averaged scores usually indicate high performance on less frequent labels, while high micro-averaged scores usually indicate high performance on more frequent labels.

5.3 Baselines

To the best of our knowledge, our work is the first study on fine-grained detection of illicit drug trafficking on Instagram. We have
Table 2: Performance comparison between the proposed method and the baseline approaches.

| Method         | Example-based | Label-based |
|----------------|---------------|-------------|
|                | Subset Acc. | Hamming loss | micro Pre | micro Recall | micro F | macro Pre | macro Recall | macro F |
| ResNet50       | 0.4187      | 0.3054       | 0.3348    | 0.7197       | 0.4570  | 0.2949    | 0.6635     | 0.3750  |
| ResNet152      | 0.4155      | 0.2844       | 0.2901    | 0.6612       | 0.4537  | 0.3453    | 0.5717     | 0.3500  |
| VGG16          | 0.4144      | 0.4066       | 0.2901    | 0.8499       | 0.4275  | 0.2855    | 0.8986     | 0.3862  |
| DenseNet121    | 0.4263      | 0.3551       | 0.2916    | 0.7559       | 0.4252  | 0.2790    | 0.7378     | 0.3576  |
| TextRNN        | 0.6775      | 0.0831       | 0.7524    | 0.6848       | 0.7170  | 0.6829    | 0.4896     | 0.5462  |
| TextCNN        | 0.6502      | 0.0956       | 0.6918    | 0.6826       | 0.6872  | 0.5853    | 0.5144     | 0.5374  |
| FastText       | 0.6604      | 0.0875       | 0.7272    | 0.6892       | 0.7077  | 0.6892    | 0.4870     | 0.5369  |
| BERT           | 0.8902      | 0.0208       | 0.9174    | 0.9711       | 0.9435  | 0.8786    | 0.9671     | 0.9185  |
| Concatenation  | 0.5727      | 0.1255       | 0.5986    | 0.9024       | 0.7197  | 0.5333    | 0.8806     | 0.6395  |
| FBC            | 0.4553      | 0.2315       | 0.4261    | 0.8547       | 0.5687  | 0.3582    | 0.8547     | 0.4655  |
| Proposed       | 0.9322      | 0.0135       | 0.9496    | 0.9765       | 0.9629  | 0.9217    | 0.9728     | 0.9455  |

Table 3: Impact of different image encoders on our method.

| Method          | Example-based | Label-based |
|-----------------|---------------|-------------|
|                 | Subset Acc. | Hamming loss | micro Pre | micro Recall | micro F | macro Pre | macro Recall | macro F |
| ResNeXt50 + BERT| 0.9107      | 0.0160       | 0.9359    | 0.9771       | 0.956  | 0.9079    | 0.9731     | 0.9386  |
| ResNeXt101 + BERT| 0.9257     | 0.0139       | 0.9458    | 0.9873       | 0.9618 | 0.914     | 0.9711     | 0.9400  |
| Vgg16+BERT      | 0.9322      | 0.0135       | 0.9496    | 0.9765       | 0.9629 | 0.9217    | 0.9728     | 0.9455  |
| Vgg19+BERT      | 0.9225      | 0.0158       | 0.9406    | 0.9729       | 0.9564 | 0.9095    | 0.9649     | 0.9434  |
| DenseNet121+BERT| 0.9182      | 0.0160       | 0.9379    | 0.9747       | 0.956  | 0.9066    | 0.9698     | 0.9358  |
| ResNet50+BERT   | 0.9311      | 0.0130       | 0.9487    | 0.9801       | 0.9641 | 0.9237    | 0.9760     | 0.9478  |
| ResNet152+BERT  | 0.9300      | 0.0142       | 0.9457    | 0.9765       | 0.9609 | 0.9152    | 0.9675     | 0.9391  |

5.4 Experimental Results

Table 2 shows the detection performances of our proposed method and the baseline approaches with different multilabel classification metrics. Our approach consistently and significantly outperforms the baseline methods on all example-based metrics and label-based metrics. We describe several key observations we have made from these results as follows:

**Text vs image in IDTE detection.** We first observe that the performance of text-only methods is better than all image-based methods. For example, even the weakest TextCNN outperforms the best image-based VGG16 by 15.12% in terms of macro F1. It indicates that textual information is more important and reliable than images for IDTE detection (contrary to the findings in [42]). We further observe that BERT model can achieve highly accurate classification performance. For example, BERT outperforms the best text-based compared our proposed method with the following baseline methods, which are modified for multilabel learning:

- **Image-only baselines.** We fine-tune pretrained CNNs (including VGG16 [37], ResNet50 and ResNet152 [15], and DenseNet121 [17]) on the images associated with IDTEs to extract the features for the multilabel classification task.

- **Text-based baselines.** We train a multilabel classification model on the textual information of IDTEs using the following models: TextRNN [44], TextCNN [24], FastText [19], and BERT [8].

- **Multimodal learning baselines.** We first use ResNet50 and BERT to extract image and text features respectively, and then apply multimodal fusion method to combine the image and text features for multilabel classification. We compare following multimodal fusion methods: concatenation [23] and factorized bilinear coding (FBC) [11, 18].

---

**Multimodal learning in IDTE detection.** The motivation of multimodal learning is that the presence of multiple information sources can be complementary to learn better feature representations for final detection of illicit drugs. However, we observe that multimodal learning does not necessarily outperform models trained with a single modality. As shown in Table 2, two popular multimodal fusion methods (i.e., concatenation and factorized bilinear coding (FBC)) are applied to composite image and text features extracted from ResNet50 and BERT respectively; but their performances are worse than single-modality based BERT model. One possible reason, as discussed in [41], is that such multimodal models are prone to overfitting when features are combined at a late stage. By contrast, building upon a powerful BERT model and projecting image embedding to the text token space, our method achieves better fine-grained multimodal fusion, thus improving the detection performances. For example, our method improves BERT by 2.7% in...
Impact of image encoders on our method. Thanks to the conceptual simplicity of MMBT, we can easily substitute different image and text encoders into the bitransformer module. Unlike text information, BERT has shown dominating performance; image encoders have shown comparable performance based on Table 2. Therefore, we have conducted an ablation study to compare different image encoders while keeping the BERT encoder the same. Table 3 includes the performance comparison among seven competing image encoders. It can be observed that on the average ResNet50 achieves the best performance, which is in contrast to the adoption of ResNet152 for MMBT in [22].

5.5 IDTE Detection Case Studies

The experimental results have shown the effectiveness of the proposed DMML approach in detecting IDTEs. We present some case studies as shown in Table 4 to demonstrate the superiority as well as limitations of the proposed method. Case 1 and Case 2 illustrate two examples of the proposed method with completely correct detection, while Cases 3-5 show the examples with missing or false detection. When compared with the results of the text-only BERT model, we can observe that by fusing text with image modality, the proposed DMML can not only effectively reduce false alarms (see Case 2 for example - LSD was incorrectly predicted by BERT but supplementary image information shows it is not LSD) in Case 3, painkiller is false alarm of BERT prediction, which gets corrected by DMML), but also detect more accurate labels (see Case 4 - BERT misses a few labels). However, for input text with special symbols that drug dealers used to disguise illicit deals, the proposed method will fail to accurately detect all labels, such as the mushroom with ‘s.H.r.o.o.M.s’ in Case 3, MDMA with ‘MdΩmaΩ’ in Case 4, and Cocaine with ‘COKE’ in Case 5. One potential solution is to design a new data preprocessing algorithm to recover these special symbols in words.

6 CONCLUSION

In this study, we have collected and constructed a large-scale multimodal IDTE dataset (MM-IDTE) from Instagram data to support the research related to illicit drug trafficking activity detection. Our dataset includes both textual and visual information contained in posted comments and has been manually annotated for multiple drug types. An automatic hashtag-based data crawling system and a user-friendly interactive web-based data annotation system were developed. Our data crawling and annotation systems allow us to build a dataset with thousands of fine-grained samples with multiple labels. Based on the constructed MM-IDTE dataset, we have developed a deep multimodal multilabel learning approach to detect suspect IDTEs and demonstrate its effectiveness on the new MM-IDTE dataset. It is shown that the proposed MMBT-based approach can dramatically outperform unimodality and ad hoc multimodal fusion strategies. We have also experimentally compared different image encoders for MMBT-based fusion and found that MMBT based on ResNet50 and BERT achieves the best performance. Extensive experimental results on the MM-IDTE dataset show that the
proposed DMML methodology can accurately detect IDTEs even in the presence of special symbols and style changes attempting to evade detection.

REFERENCES

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. 2017. Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017).

[2] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multi-modal machine learning: A survey and taxonomy. TPIR 41, 2 (2020), 423–443.

[3] Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. 2015. Sparse local embeddings for extreme multi-label classification. In NIPS. 730–738.

[4] Cody Buntin and Jennifer Golbeck. 2015. This is your Twitter on drugs: Any questions? In WWWW. 777–782.

[5] Jonathan P Cauklin, Beau Kilmer, and Mark AR Kleiman. 2016. Marijuana legalization: What everyone needs to know. Oxford University Press.

[6] Edward Chang, Kingshly Goh, Gerard Schayy, and Gang Wu. 2003. CBSSA: context-based soft annotation for multimodal image retrieval using Bayes point machines.

[7] TCSVT 13, 1 (2009), 26–38.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

[9] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016).

[10] Yuqi Gao, Jiaxiang Sang, Tongwei Ren, and Changsheng Xu. 2017. Hashtag-centric evade detection. arXiv preprint arXiv:1607.01759 (2016).

[11] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2018. Multimodal machine learning: A survey and taxonomy. In EMNLP. 1746–1751.

[12] Ryan Kiros, Yazhe Qin, Yasunori Yamaguchi, and Rie Johnson. 2015. Skip-thought vectors. arXiv preprint arXiv:1506.06726 (2015).

[13] Zhenzhong Lan, Mengda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A healthy bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019).

[14] Jiawei Li, Qin Xu, Neal Shah, and Tim K Mackey. 2019. A machine learning approach for the detection and characterization of illicit drug dealers on Instagram: model evaluation study. JMR 21, 6 (2019), e13803.

[15] Changzheng Li, Wenyi Li, Baoqin Liu, and Jun Wei. 2018. Deep learning for extreme multi-label text classification. In SIGIR. 115–124.

[16] Jiajun Lu, Dhiru Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visuolinguistic representations for vision-and-language tasks. arXiv preprint arXiv:1908.62265 (2019).

[17] A Deep multi-modal CNN for large-scale multi-instance image classification. TIP 27, 12 (2018), 6025–6038.

[18] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In CVPR. 806–813.

[19] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).

[20] Lingyong Song, Jun Liu, Buyue Qian, Mingxuan Sun, Kuan Yang, Meng Sun, and Samaar Abbas. 2018. A deep multi-modal CNN for multi-instance multi-label image classification. TIP 27, 12 (2018), 1079–1089.

[21] Yiwei Wang, Wu Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard?. In CVPR. 12705–12705.

[22] Xiaotong Yang and Jiebo Luo. 2017. Tracking illicit drug dealing and abuse on Instagram using multimodal analysis. TIST 8, 4 (2017), 1–15.

[23] Zhihan Yang, Zhigang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237 (2019).

[24] Pengpeng Yin, Kvin Liu, Adam J Console, and Huajun Jiang. 2018. Solution to detect, classify, and report illicit online marketing and sales of controlled substances via Twitter: Using machine learning and web forensics to combat digital opioid access. JMIR 20, 4 (2018), e10029.

[25] Tomas Mikolov, Bya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1301.3781 (2013).

[26] Maxime Oquab, Leon Bottou, Ivar Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In CVPR. 1717–1724.

[27] Sarah J Reddy, Satyen Kale, and Sanju Kumar. 2019. On the convergence of adam and beyond. arXiv preprint arXiv:1904.09257 (2019).

[28] Abeer Sarker, Annika DeReos, and Jeanmarie Perrone. 2020. Mining social media for prescription medication abuse monitoring; a review and proposal for a data-centric framework. JAMIA 27, 2 (2020), 315–329.

[29] Abeer Sarker, Graciela Gonzalez-Hernandez, Yucheng Ruan, and Jeanmarie Perrone. 2019. Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter. JAMIA open network 2, 11 (2019), e1914672–e1914672.

[30] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In CVPR. 806–813.

[31] Shunfeng Zhu. 2019. Attentionxml: Label tree-based attention-aware deep model for large-scale multi-label image classification. TKDE 25, 7 (2010), 1079–1089.

[32] Fei-Fei Liu, Soon A Chun, James Geller, Huy Vo, Xinyue Ye, Ruoming. 2018. Solution to detect, classify, and report illicit online marketing and sales of controlled substances via Twitter: Using machine learning and web forensics to combat digital opioid access. JMIR 20, 4 (2018), e10029.

[33] Tomas Mikolov, Bya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1301.3781 (2013).

[34] Maxime Oquab, Leon Bottou, Ivar Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In CVPR. 1717–1724.

[35] Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. 2018. Binary relevance multi-label learning. In Proc. 20th ACM SIGKDD int’l conf. knowl. discovery and data mining. 861–869.

[36] Jon Shaposhnik, Khosrow Homayoun, Xin Zhao, and Zhihua Zhou. 2013. A review on multi-label learning methods and approaches to combat prescription drug abuse. Current Addiction Reports 4, 4 (2017), 397–409.

[37] Yiheng Zhou, Numair Sani, Chia-Kuei Lee, and Jiebo Luo. 2016. Understanding task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv preprint arXiv:1904.09257 (2019).

[38] Chengzhong Zhou, Ruiyuan Chen, and Jiebo Luo. 2017. Tracking illicit drug dealing and abuse on Instagram using multimodal analysis. TIST 8, 4 (2017), 1–15.

[39] Zhihan Yang, Zhigang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237 (2019).

[40] Pengpeng Yin, Kvin Liu, Adam J Console, and Huajun Jiang. 2018. Solution to detect, classify, and report illicit online marketing and sales of controlled substances via Twitter: Using machine learning and web forensics to combat digital opioid access. JMIR 20, 4 (2018), e10029.

[41] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1301.3781 (2013).

[42] Maxime Oquab, Leon Bottou, Ivar Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In CVPR. 1717–1724.

[43] Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. 2018. Binary relevance multi-label learning. In Proc. 20th ACM SIGKDD int’l conf. knowl. discovery and data mining. 861–869.

[44] Jon Shaposhnik, Khosrow Homayoun, Xin Zhao, and Zhihua Zhou. 2013. A review on multi-label learning methods and approaches to combat prescription drug abuse. Current Addiction Reports 4, 4 (2017), 397–409.