Making Sense of Darknet Markets: Automatic Inference of Semantic Classifications from Unconventional Multimedia Datasets

Alexander Berman[1] and Celeste Lyn Paul[2]

[1] Texas A&M University. anberman@tamu.edu
[2] U.S. Department of Defense. clpaul@tycho.ncsc.mil

Abstract. Darknet Markets are a hotbed of illicit trade and are difficult for law enforcement to monitor and analyze. Topic Modeling has been a popular method to semantically analyze market listings, but lacks the ability to infer the information-rich visual semantics of images embedded within these listings. In this paper we present a relatively fast method using unsupervised and self-supervised machine learning methods to infer image semantics from large, unstructured multimedia corpora, and demonstrate how it may aid analysts in investigating the content of Darknet Markets.

1 Introduction

Darknet Markets (DNM) are online marketplaces hosted on anonymous networks, such as Tor, that provide access to buyers and sellers of usually illicit or unregulated goods. A famous DNM, Silk Road, operated from 2011 until 2013 when it was shut down by the FBI for illegal trade [1, 12]. Silk Road was the first of dozens of DNMs to operate - and be shut down by law enforcement. Law enforcement is in a constant game of whack-a-mole as two new markets pop up for every one that is taken down.

There is an increasing need to understand the scope and content of DNMs to support law enforcements efforts to protect public safety. As markets are shut down, new markets will often adopt the abandoned content, making tracking overall trends an essential task. The sheer scope and quantity of DNMs makes this a significant big data analysis problem. Common approaches to this problem include textual content analysis [12, 17]. However, these methods fall short of fully characterizing DNM content since listings tend to include multimedia, such as images, to describe the products rather than using solely text. The content itself can be a challenge to parse consistently for use in Natural Language Processing (NLP) analysis [17], requiring manual and time-consuming methods as a fall-back.

Methods that provide automatic semantic organization of DNM listings based on multimedia properties could help make the analysis of these markets more tractable. There has been some success utilizing supervised machine learning algorithms to classify DNM images through training on a large labeled corpus of images from market listings [15]. However, labeling large image corpora and training of new image classifiers is a non-trivial task. To properly label these datasets, analysts must spend many hours distinguishing the semantics between all images in a dataset, or analysts must find an existing labeled dataset that is analogous to their goals. In the case of DNMs, there are few datasets containing labels for illicit object images, such as specific guns and drugs. The labeling process can also be prone to human bias and issues of not sampling less frequent classes or not sampling emerging classes. Reducing the labeling burden on analysts and could allow for more timely and accurate analysis of DNM content.

Other work on analyzing DNMs has utilized tools such as topic modeling to help track activities within these markets [6, 12], and bag-of-visual-words methods to classify images [6]. This paper presents a method to automatically identify semantic relationships between text and images for listings in the Gwern Darknet Markets dataset [9], in order to aid analysts in achieving a better understanding of diverse multimedia descriptions from various illicit goods and services. We demonstrate applications of this method on DNM listings, and propose future work to assist with future DNM content analysis.

2 Related Work

Image classification systems with near-human accuracy [23] have increased interest for organizing and inferring classifications of images across many domains. Previous work has taken the approach of labeling large datasets for the purpose of classifying and tracking DNM activity. Fidalgo et al. categorized Darknet images that corresponded with particular illegal activities [6]. They utilized Edge-SIFT features with dense SIFT descriptors to categorize pages on the Darknet by images with high accuracy. This is particularly applicable to law-enforcement and security agencies, which could benefit from being able to automatically flag and sort through illegal Darknet activities.

As new forms of illegal activity become prominent, there is a need from law enforcement to be able to classify those activities promptly. However, these classification systems often require large amounts of labeled data. In many domains, such as identifying product images in niche online markets, it takes large-scale crowd-sourcing [3], or smaller groups of people long periods of time to label enough images to train classification systems to satisfactory performance. In addition, these labeled images may not meet every future analysis need. The only semantic-inference that can be done with the supervised models trained on the labeled dataset is from the pre-defined semantics designated by the labelers. This means that it can take several iterations of labeling entire datasets before being able to predict future images for certain analysis tasks. Also, this predicates that several individuals need to spend non- trivial amounts of time looking through images and determining distinguishable semantic categories of images before labeling each image.

For many applications, the time, specialty, and clearance needed to properly label a dataset may be infeasible. Where possible, automatic labeling of images based on an images surrounding context could make classification feasible. To accomplish this, descriptive labels from other media such as text must be generated to describe this context. In Porter's analysis of Darknet market terms over time [12], he utilized Latent Dirichlet Allocation (LDA) on text data scraped from Reddit forums that discussed darknet activities. Porter generates high-level topics of these subreddits, and identifies changes in behavior after many darknet markets faced legal action in summer of 2017[12]. A notable aspect of this change was a shift from casual language to more serious, security-concerned language. He notes that while LDA could give great insight into changes within the darknet community, all insights should be verified with the original textual source. Porter's LDA-driven analysis was much faster than analyzing the text from scratch, as one can form hypotheses quicker; but, the analysis still takes some time. Our work addresses some of these issues by augmenting the topic model latent space with correlated images, which can help inform analysts and further support their hypothesis formation.

Another method for saving time in analysis of webpages is TextTopicNet [7, 19], which presents a method to bootstrap image classification and cross-modal retrieval tools by training an instance of CaffeNet [13] to predict semantic meaning of images, optimizing to match images to their source document topic vectors. Using the ImageCLEF Wikipedia dataset [24], Gomez et al. demonstrate retrieval based on LDA topic vector weights, and on the learned visual features from the Convolutional Neural Network (CNN). They show that training CNNs to match broad semantic topics can better support more specific computer vision tasks like image classification, object detection, and multi-modal retrieval. However, the time and expertise needed to train TextTopicNet would be presently infeasible by an analyst not well-versed in neural networks, and without access to a GPU-accelerated workstation to speed up training. Inspired by TexTopicNet, we utilize a similar method to automatically train a model within a more feasible timeframe, and demonstrate applications of its image topic-composition predictions on Wikipedia articles and Darknet marketplaces.

3 Methodology

Our method for supporting the analysis of DNMs is inspired by previous work in the area. First, similar to TextTopicNet [7, 19], we created training image-topic vectors (labels) for each image from the LDA-generated document-topic vectors from where that image originated. Then, we trained a convolutional neural network (CNN), accelerated by transfer-learning [25], to predict the document topic-vector where a given image could be located. Once the training converged, a matrix of all images in the dataset and the predicted textual topic vectors was created. This model essentially predicts the type of document from which a given image would originate, affording analysts the ability to directly relate images and text in the domain of the training dataset.

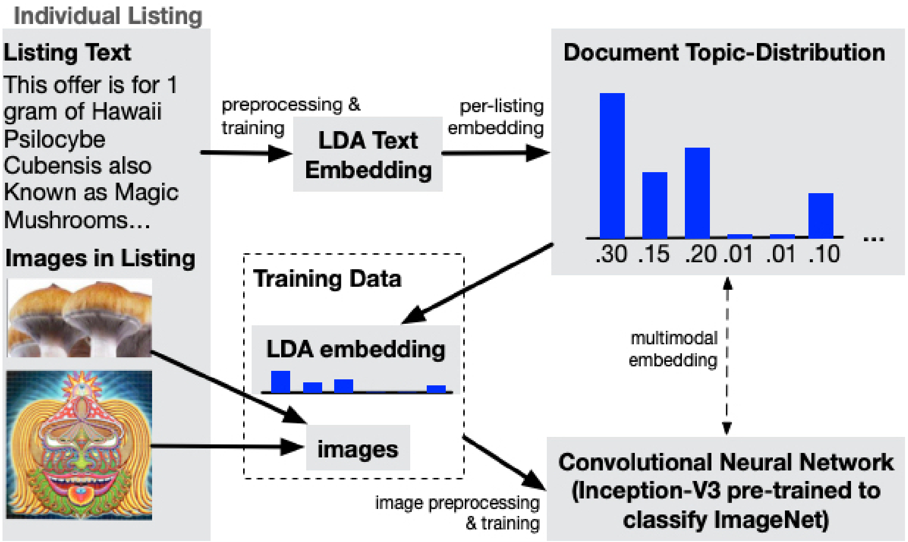

Individual Listing
Listing Text
This offer is for 1 gram of Hawaii Psilocybe Cubensis also Known as Magic Mushrooms...
Images in Listing
preprocessing & training
LDA Text Embedding
per-listing embedding
Document Topic-Distribution
.30 .15 .20 .01 .01 .10 ...
Training Data
LDA embedding
images
multimodal embedding
image preprocessing & training
Convolutional Neural Network (Inception-V3 pre-trained to classify ImageNet)

Fig. 1. System for training a CNN to predict Topic Vectors for any given Image, representing what type of text document an image with similar visual/semantic features would be found according to this model.

3.1 Latent Dirichlet Allocation

In order to automatically classify images based on associated documents, we make the assumption that images are correlated to probability distributions of topics that generate text documents within the dataset. These compositions of textual topics are generated via LDA topic modeling. LDA is a generative statistical model of a large text collection, where each document is generated as a mixture of k topics (where k is specified by the user). Each topic is represented as a probability distribution over words present in the text collection. The result of this generative process, as defined by Blei et al.[2], is two parameters: word probabilities given each topic, and topic probabilities given each document. All documents in the dataset are represented by topic probabilities, which are in turn represented by word probabilities. Any document, even if unseen during the training of an LDA model, can be represented by a probability distribution over all topics of the learned model. We utilize the MALLET library to train LDA models and tune its hyper-parameters [16].

Representing text and predicting images in topic space, instead of Bag-of-Words representations, provides semantically meaningful descriptors in lower-dimensional space. By representing images in the same latent semantic space as text, we can more comparably analyze how textual semantic features and image visual features relate to one-another within a corpus. This allows analysts to search for images that would co-occur in with other images, images that would likely occur with specified text, documents that would likely accompany images, and documents that are similar to a specified text (see Figure 2).

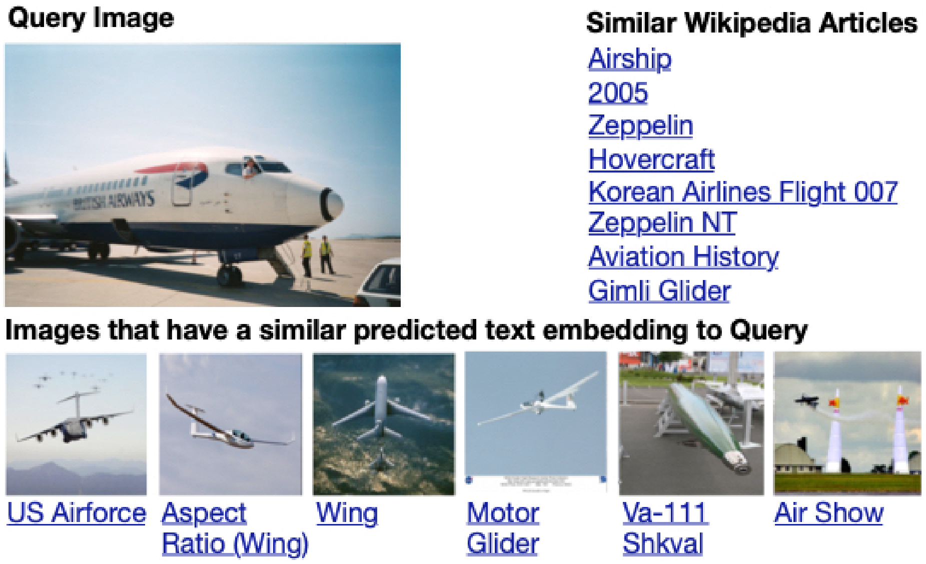


Fig. 2. The CNN trained in this paper can quickly retrieve documents that a query image may be found within, and identify other images that may be found in those types of documents. The above example demonstrates retreiving Wikipedia articles and images from the ImageCLEF dataset with a 200-topic LDA and CNN model.

3.2 Neural Network Architecture and Training Procedure

To automatically create categories of images within a textual dataset without much human supervision, we adapted TextTopicNet [7, 19] to work on Inception-V3 image classification CNN architecture [23]. TextTopicNet took 120,000 epochs (the number of times it trained on the entire image dataset) to converge on the ImageCLEF dataset [24]. This would take an infeasible amount of time on many everyday machines, and still would take a non-trivial amount of time on multi-GPU machines. To combat this, we applied transfer-learning from a model trained on image-classifications from the ImageNet dataset, which contains over 1000 different classifications for everyday images [5]. On top of the convolutional layers of Inception V3 [23], we placed a global average pooling layer, followed by a fully connected layer with 1024 output dimensions, a dropout layer with 0.5 probability, and another fully connected layer that outputs dimensions equal to the k dimensions of the trained LDA model. This network is then optimized with Stochastic Gradient Descent of learning rate of 0.0001 and momentum of 0.9, to minimize sigmoid cross-entropy loss similar to TextTopicNet.

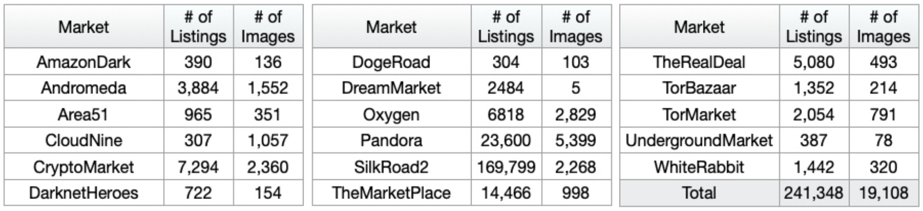
Once an LDA model is trained on a textual dataset, we created a table with rows corresponding to the images, text, and the topic vectors. This is randomly split into a training dataset (80%) and a validation dataset (20%). Training is done in two stages: an initial transfer-learning stage, and an additional fine-tuning stage. The first stage loads the ImageNet-trained weights into the model, initializing the fully-connected layers with variance scaling [11]. All layers besides the dense layers are frozen, so that the dense layers train while utilizing the image-features recognized by the ImageNet-trained convolutional layers. After the validation loss converges or a maximum number of epochs are reached, this stage ends. The set of CNN weights with the best epoch performance are loaded for the fine-tuning stage. The top 172 convolutional layers are then unfrozen and updated by the next round of training. This allows for the higher-level features generated by the convolutional network to be better fit to the training images, which may require different features than ImageNet. Once this stage converges or reaches a maximum number of epochs, training is stopped. The weights with the best overall validation-loss are loaded and then topic vectors are generated for each image. For 40-topic LDA models, we trained with an upper-limit of 25 epochs per the transfer and fine-tuning stages seperately (50 combined), ending each stage if validation accuracy increased for three successive epochs. For LDA models with larger topic-dimensions, the upper-limit was increased to 50 epochs. The neural network components of this system were implemented using the Keras python framework [4]. With transfer-learning it typically takes fewer than 40 epochs total to converge rather than the more than one-hundred-thousand epochs used in the original TextTopicNet model [19, 7]. This allows for relatively fast training, that could make this method affordable and accessible for future applications in analysis of multimodal data, such as DNM.

3.3 Datasets

To test this neural network training process and to investigate applications of the trained model, we trained on two datasets: ImageCLEF and Gwern Darknet Markets. To provide evidence that our system works similarly to TextTopicNet [7, 19], we train and provide results on models from the ImageCLEF dataset [24] which serves as a common link between our papers. We then train models and demonstrate results on Darknet listings to convey this system's potential to aid analysts in navigating more unconventional multimodal datasets. ImageCLEF is a collection of Wikipedia articles with images from 2010-2011 [24]. The Gwern Darknet Markets Archive includes HTML documents of DNM listings and includes images [10]. We utilized a subset of 17 markets from this dataset in all training and demonstrations for this paper.

The ImageCLEF dataset [24], same as the one utilized by TextTopicNet [7, 19], has 42,777 English Wikipedia articles between 2010-2011 with 100,785 associated images. All articles have been completely cleaned of HTML artifacts, and only contain article-specific text. Articles and images cover a large semantic range. This paper reproduces those results on the recommended 40 topics, as well as increasing number of topics beyond mentioned in TextTopicNet. A standard

English stopword list was applied to this dataset before performing LDA. This dataset was utilized to test our methodology before utilizing the system on the less-studied DNM.



Market	# of Listings	# of Images
AmazonDark	390	136
Andromeda	3,884	1,552
Area51	965	351
CloudNine	307	1,057
CryptoMarket	7,294	2,360
DarknetHeroes	722	154
DogeRoad	304	103
DreamMarket	2484	5
Oxygen	6818	2,829
Pandora	23,600	5,399
SilkRoad2	169,799	2,268
TheMarketPlace	14,466	998
TheRealDeal	5,080	493
TorBazaar	1,352	214
TorMarket	2,054	791
UndergroundMarket	387	78
WhiteRabbit	1,442	320
Total	241,348	19,108

Fig. 3. Number of Listings and Images analyzed from the Gwern Darknet Markets Archive per market. An image may be referenced in multiple listings within a market.

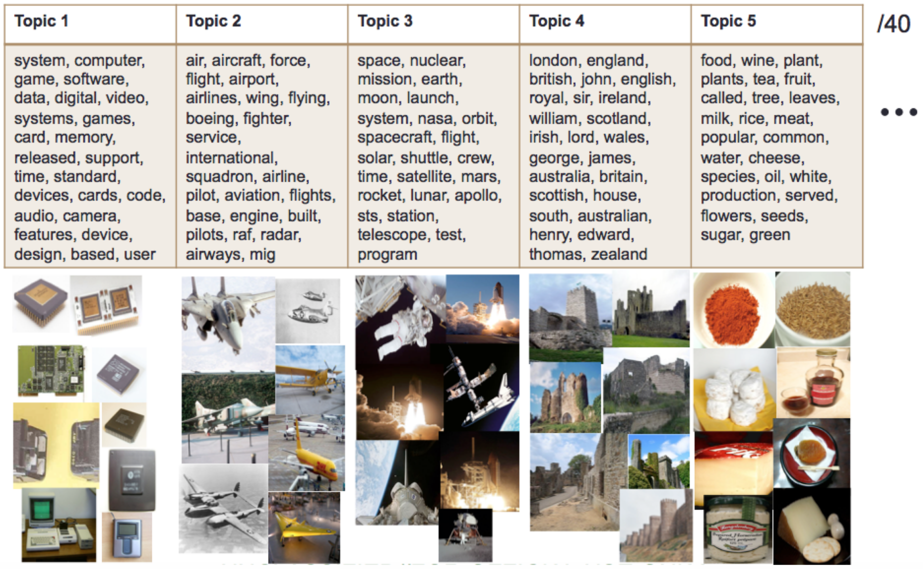
The Gwern Darknet Markets Archive is a publicly-available dataset gathered by crawling various DNMs [10]. In this paper, we analyze 17 markets from this dataset (see Figure 3). These markets consist of raw HTML from market listings across various dates. We only processed the latest version of duplicate listings from the crawlers datasets, ignoring older listings with less text and images. Then we removed HTML tags from the listings. As many artifacts and uninformative text remained in the listings, the top 20% most-frequent terms per market were removed from the respective DNM listings. This removed many missed webpage artifacts, and some repeated other text such as headers and market-wide text. Both images downloaded in the crawler, and embedded images from the HTML were associated with listings. We generated a stoplist of commonly missed HTML artifacts and less-informative terms. Researchers removed the most common images per market that had nothing to do with the listings (e.g. button icons, market banners, etc.). Small images were also removed. A total of 241,348 listings were extracted with 47,101 images. LDA and the CNN were trained with 40 topics and 450 topics.

4 Results

We will first describe what the image textual-topic prediction models can accomplish with a clean dataset like Wikipedia, before moving to the models trained on the noisier Darknet dataset. Insights gained from the Wikipedia data were then applied to the Darknet dataset to help find patterns without much prior knowledge of the markets.

4.1 ImageCLEF

We trained a image to textual-topic prediction model using the method described in 3.2 of the Methodology on 40 topics. A sample of the top-weighted



Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
system, computer, game, software, data, digital, video, systems, games, card, memory, released, support, time, standard, devices, cards, code, audio, camera, features, device, design, based, user	air, aircraft, force, flight, airport, airlines, wing, flying, boeing, fighter, service, international, squadron, airline, pilot, aviation, flights, base, engine, built, pilots, raf, radar, airways, mig	space, nuclear, mission, earth, moon, launch, system, nasa, orbit, spacecraft, flight, solar, shuttle, crew, time, satellite, mars, rocket, lunar, apollo, sts, station, telescope, test, program	london, england, british, john, english, royal, sir, ireland, william, scotland, irish, lord, wales, george, james, australia, britain, scottish, house, south, australian, henry, edward, thomas, zealand	food, wine, plant, plants, tea, fruit, called, tree, leaves, milk, rice, meat, popular, common, water, cheese, species, oil, white, production, served, flowers, seeds, sugar, green

Fig. 4. 5 out of 40 Topics from the ImageCLEF Wikipedia dataset with Top Terms and Top Images by Individual Weight per Topic. Wikipedia articles are a mixture of these topics terms and image characteristics

images per topic is shown in Figure 4. These resemble many of the results from TextTopicNet [7, 19], which was trained on the same Wikipedia dataset. These similar results validate that our differences from TextTopicNet, namely the InceptionV3 CNN architecture [23] and transfer-learning [18], did not noticeably alter the end results for how the model predicts LDA textual topic space that would contain any given image. Many of the images, in each topic, show a mixture of many objects fitting a topical theme. Airplane categories (Figure 4, Topic 2) include everything from commercial airlines to fighter jets. A topic about animals has images of pets, fish, and dinosaurs. The food topic (Figure 4, Topic 5) has everything from soup to nuts. The image-classifier is able to properly summarize images belonging to different textual topics, showing dataset-specific semantic similarities between images instead of distinguishing images based on purely visual features or only strict supervised labeling patterns.

While the TextTopicNet demonstrated their network on at most one-hundred LDA topics[7, 19], in many practical applications of LDA the number of topics will be significantly larger. Many topics in the Wikipedia 40-topic LDA model contained many different classifications of images, while fitting semantic themes generated by LDA, hinted that there may be more granular themes that LDA could discover within the dataset. The square root of the number of documents has been recommended as a good number of topics for training LDA models [8], therefore we also trained LDA and the CNN with 200 topics to see the effectiveness of this type of self-supervised network with more granular topics.

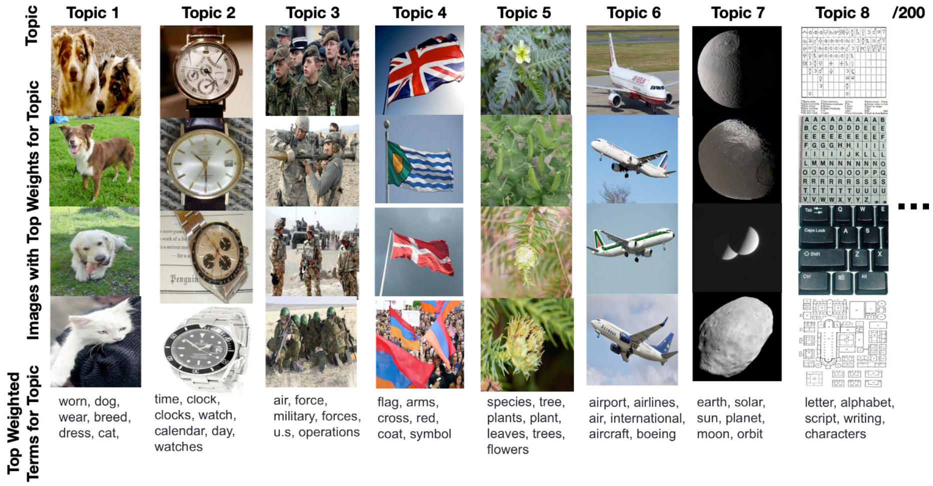

Topic
Images with Top Weights for Topic
Top Weighted Terms for Topic
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8
/200
worn, dog, wear, breed, dress, cat,
time, clock, clocks, watch, calendar, day, watches
air, force, military, forces, u.s, operations
flag, arms, cross, red, coat, symbol
species, tree, plants, plant, leaves, trees, flowers
airport, airlines, air, international, aircraft, boeing
earth, solar, sun, planet, moon, orbit
letter, alphabet, script, writing, characters

Fig. 5. 5 out of 200 Topics from the ImageCLEF Wikipedia dataset with Top Terms and Top Images by Individual Weight per Topic. Both topics and images are more specific than the 40-topic model.

The 200 topic model resulted in many more specific topics (see Figure 5). Instead of animals topic, there were topics with images showing only dinosaurs, pets, and fish separately (e.g. Figure 5, Topic 1). Similarly airplanes had multiple topics, with images in groups of only commercial planes or military planes (e.g. Figure 5, Topic 6). New topics emerged, such topics containing only wrist-watch images and images of flags (Figure 5, Topics 2 and 6). Some diagram images were more evenly split, such as a topic showing all images with grids of alphabet characters (Figure 5, Topic 8). This trained model allows for fast association between images, terms, and source documents. Some specific topics may not have a clear or obvious theme with associated images (e.g. Figure 6). While some of these themes do not seem useful at first, they can still provide useful information to analysts. For example, A topic with words describing communist revolutions had mainly black and white photos, which may relate to propaganda. A topic with words related to various media (e.g. radio, align, television, center, news, broadcast, channel, etc.) showed images of fountains, buildings and concerts. These topics show a clear pattern between text and related images, while it may not be obvious without extra thought of why they connect. Other predicted image-text relations may provide evidence of lack of clear themes in the corpus. For example, a topic of people's names has diverse images associated with it, and a topic of scientific terms has many science-related pictures without a clear theme. When inspecting the Wikipedia articles these images originate from, they do contain terms from these topics. Not all topics will have clear patterns associated with the images, especially at higher topic dimensions, but this also can provide useful information to an analyst. Models with lower topic-dimensionality can provide analysts with more interpretable text-image relations, while models with higher

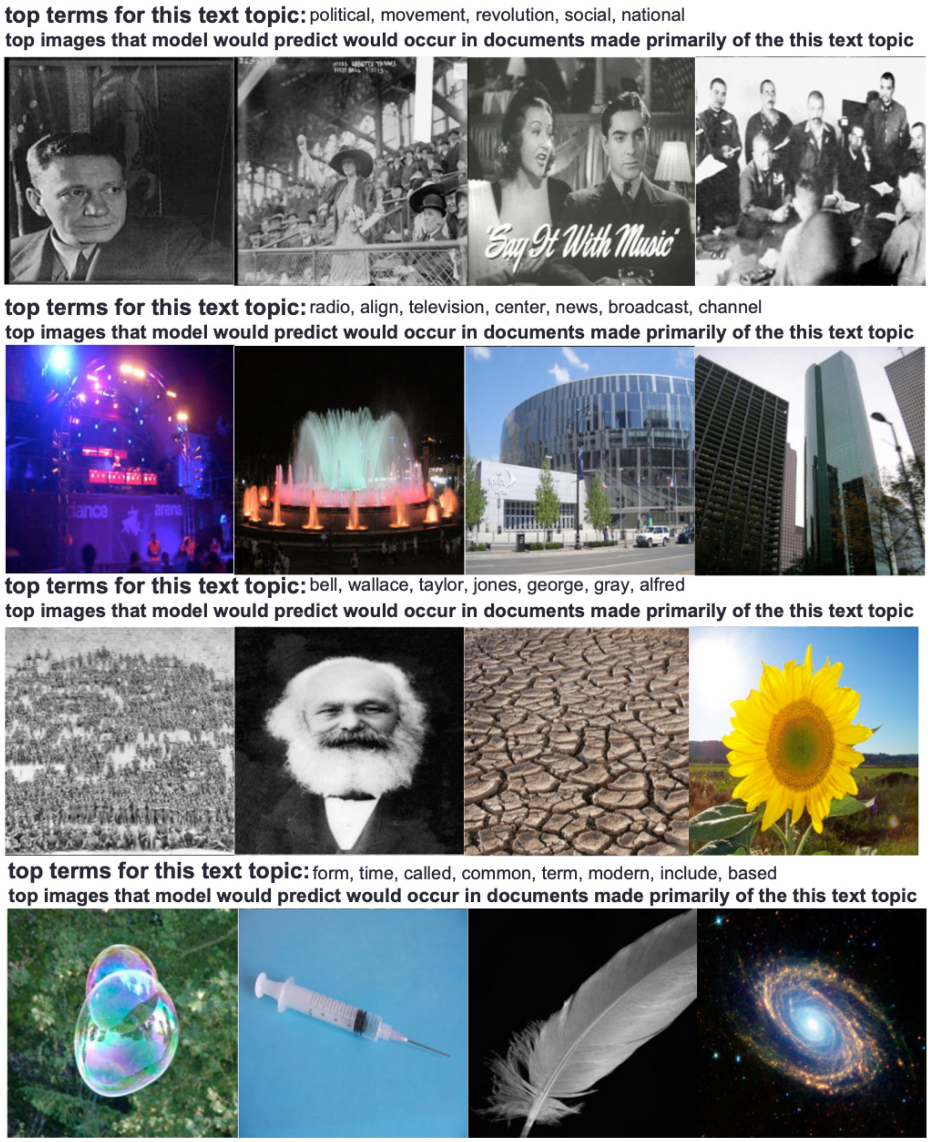

top terms for this text topic: political, movement, revolution, social, national
top images that model would predict would occur in documents made primarily of the this text topic
'Say It With Music'
top terms for this text topic: radio, align, television, center, news, broadcast, channel
top images that model would predict would occur in documents made primarily of the this text topic
dance
arena
top terms for this text topic: bell, wallace, taylor, jones, george, gray, alfred
top images that model would predict would occur in documents made primarily of the this text topic
top terms for this text topic: form, time, called, common, term, modern, include, based
top images that model would predict would occur in documents made primarily of the this text topic

Fig. 6. Top weighted terms and images for topics in the 200-topic ImageCLEF model. Images associated with topics can provide meaningful insight into related source documents, even when that association is not always obvious.

topic-dimensionality will provide more precise relations between text-image relations that may be less interpretable. Both models have advantages, so analysts may desire to utilize both in conjunction to maximize interpretability while still achieving specificity.

Image nearest-neighbor results also leads to more semantically specific image results than when run on fewer topic dimensions. An image of Aaron Rodgers throwing a football was close via Jensen-Shannon Divergence [14] to not only football action-shots, but also an image of the University of Oklahomas mascot. The closest similarities to a given image or text query constitute a theme, not an exact matching to the query. Analysts utilizing this system will have to find an optimal number of topics, which has been an open question in LDA research [8]. Image summary and retrieval tasks primarily inform semantics based on the granularity and quality of the topic model, meaning the LDA model quality is imperative.

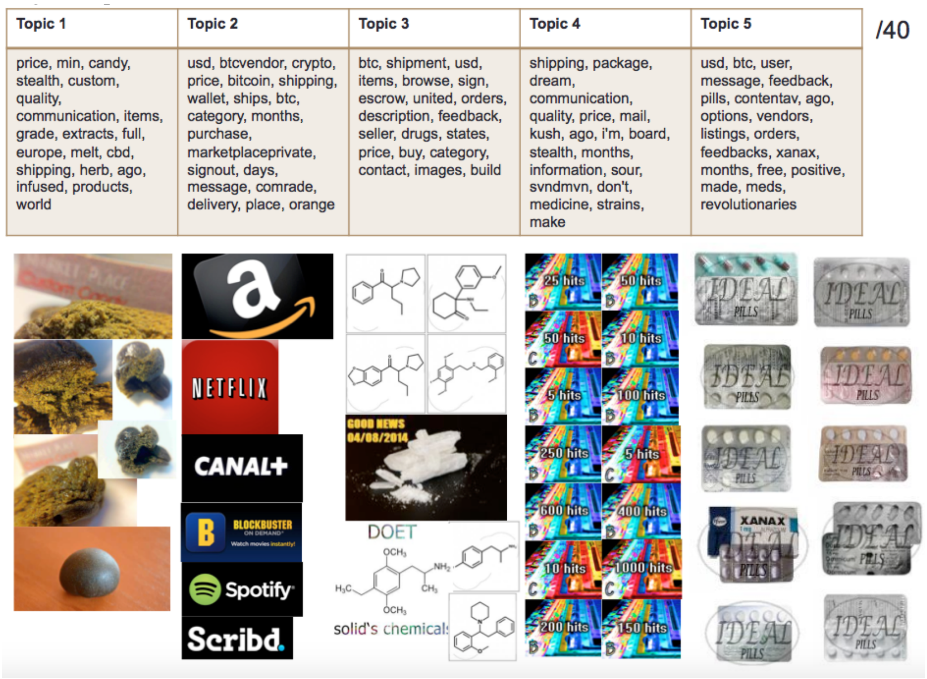
4.2 Gwern Darknet Markets Archives

The Darknet Markets Archive [10] was a much messier dataset than the cleaned ImageCLEF Wikipedia articles, and there were a number of artifacts not captured by our preproccessing that appeared in the topic space, such as some HTML artifacts, common artifacts from markets (e.g., countries where they can ship), multi-language descriptions, and less polished and consistent wording in many listings. The difficulty in interpretability of the top words per topic does make understanding the model more difficult, but the generated groups of images can still deliver insight into the DNM listings. In these ways, the DNM dataset is more representative of many less-structured datasets than the curated ImageCLEF.

Images Inform Context of Text Topics Despite the extra artifacts, clear image themes emerge from this dataset, in the 40-topic and 450-topic (approximately the square root of the total number of listings) LDA models. From identifying semantic patterns between images and textual topics, an analyst can gain a sense of the types of images that are in different types of DNM listings. Through this process, viewing images associated with textual topics can aid an analyst in understanding multimodal datasets where solely text would deliver incomplete or noisy information.

Training the image model from a 40-topic trained LDA model produces many interesting high-level themes that are not obvious from looking at only the weighted terms per topic. For example (shown in Figure 7, Topic 2) a topic focused on currencies and shipping information has images of different online audio and video streaming service logos. Some images relate to details of the text, such as images showing “<number> hits” with a letter grade in the bottom-left corner that all very accurately describe a topic concerned with shipping, prices, and quality of drugs.

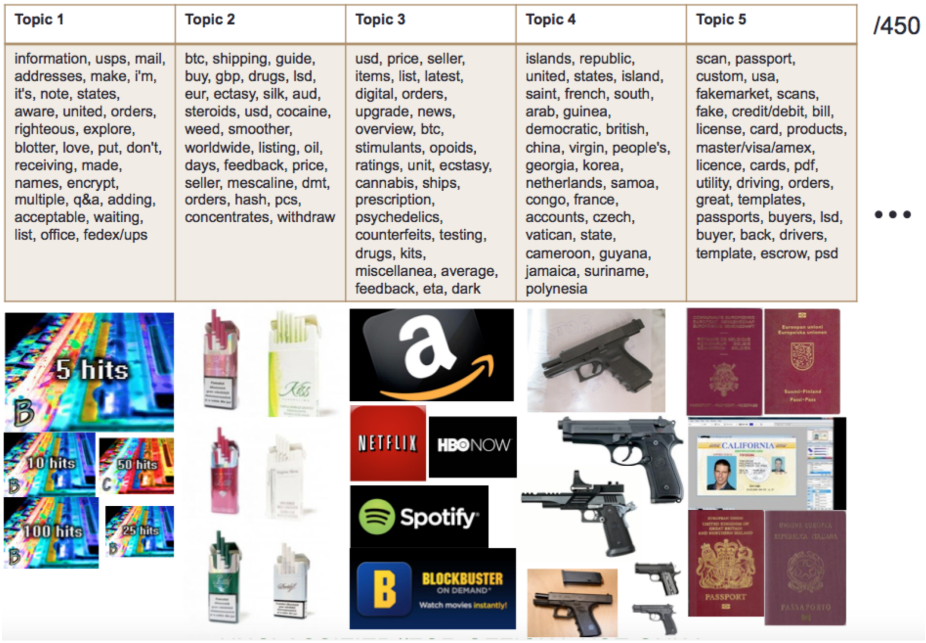
Some topics have a mixture of easily distinguishable items, similar to the mixture of terms that compose a topic. One topic has terms guide, make, and



Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
price, min, candy, stealth, custom, quality, communication, items, grade, extracts, full, europe, melt, cbd, shipping, herb, ago, infused, products, world	usd, btcvendor, crypto, price, bitcoin, shipping, wallet, ships, btc, category, months, purchase, marketplaceprivate, signout, days, message, comrade, delivery, place, orange	btc, shipment, usd, items, browse, sign, escrow, united, orders, description, feedback, seller, drugs, states, price, buy, category, contact, images, build	shipping, package, dream, communication, quality, price, mail, kush, ago, i'm, board, stealth, months, information, sour, svndmvn, don't, medicine, strains, make	usd, btc, user, message, feedback, pills, contentav, ago, options, vendors, listings, orders, feedbacks, xanax, months, free, positive, made, meds, revolutionaries

Fig. 7. 5 out of 40 Topics trained on DNM listings

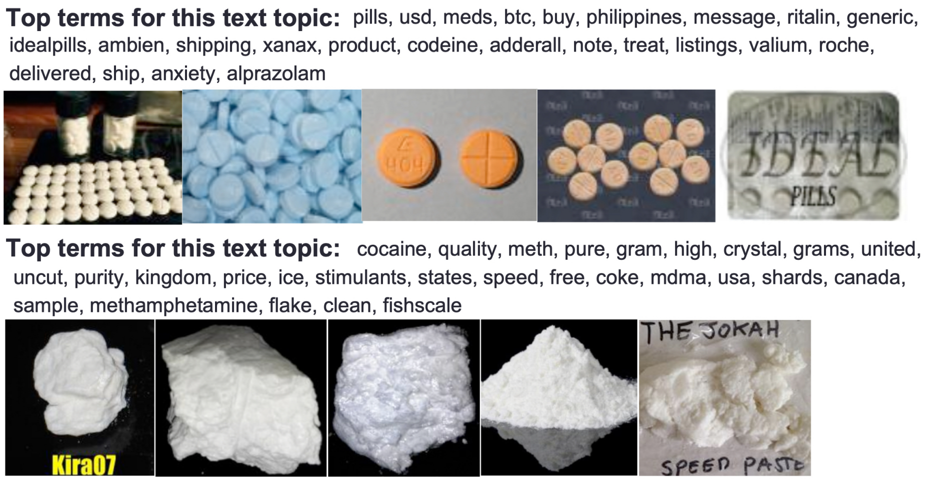
watch which is difficult to make what exactly it is talking about alone. With images, which show images of books and watches, we can infer that this topic is about both books and watches separately (Figure 10).This likely means similar wording is employed to sell watches and books on DNMs. Another example of this is a mushroom-selling topic, which has many more psychedelic images and some images related to money-exchange. The image-inference model can augment analysts understanding of topics beyond what is represented by words alone.



Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
information, usps, mail, addresses, make, i'm, it's, note, states, aware, united, orders, righteous, explore, blotter, love, put, don't, receiving, made, names, encrypt, multiple, q&a, adding, acceptable, waiting, list, office, fedex/ups	btc, shipping, guide, buy, gbp, drugs, lsd, eur, ectasy, silk, aud, steroids, usd, cocaine, weed, smoother, worldwide, listing, oil, days, feedback, price, seller, mescaline, dmt, orders, hash, pcs, concentrates, withdraw	usd, price, seller, items, list, latest, digital, orders, upgrade, news, overview, btc, stimulants, opoids, ratings, unit, ecstasy, cannabis, ships, prescription, psychedelics, counterfeits, testing, drugs, kits, miscellanea, average, feedback, eta, dark	islands, republic, united, states, island, saint, french, south, arab, guinea, democratic, british, china, virgin, people's, georgia, korea, netherlands, samoa, congo, france, accounts, czech, vatican, state, cameroon, guyana, jamaica, suriname, polynesia	scan, passport, custom, usa, fakemarket, scans, fake, credit/debit, bill, license, card, products, master/visa/amex, licence, cards, pdf, utility, driving, orders, great, templates, passports, buyers, lsd, buyer, back, drivers, template, escrow, psd

Fig. 8. 5 out of 450 Topics trained on DNMs listings

For the LDA model with 450 topics, the image to textual topic prediction model has many more specific topics (Figure 8).Individual drugs are now sorted into their own topics (Figure 9), with topics solely containing images of cocaine, prescription pills, and cigarette variations (Figure 8, Topic 2). Books gain their own topic, largely associated with hacking terms and the “blackhand. Wristwatches also get their own topic, now separated from books (e.g. Figure 10). While many of these topics have terms match to images with similar semantic meaning, some images and terms fit semantic themes on their own but not together. For example, a topic showing all images of guns relates to a topic composed of country names (Figure 8, Topic 4). This likely is due to the inherent messiness of the dataset, but still shows a correlation between terms and semantic content that otherwise would not have been obvious to an analyst viewing



Top terms for this text topic: pills, usd, meds, btc, buy, philippines, message, ritalin, generic, idealpills, ambien, shipping, xanax, product, codeine, adderall, note, treat, listings, valium, roche, delivered, ship, anxiety, alprazolam

Top terms for this text topic: cocaine, quality, meth, pure, gram, high, crystal, grams, united, uncut, purity, kingdom, price, ice, stimulants, states, speed, free, coke, mdma, usa, shards, canada, sample, methamphetamine, flake, clean, fishscale

Fig. 9. Drug Images Distinguished via CNN-Predicted association with 450 Textual Topics, allowing analysts the ability to identify items like drugs via images and text.

the noisy text data on its own. Listings with gun images, for mostly selling purposes, are likely to be on pages that list many different countries. The objective of this LDA-image model is not to link images with semantic labels describing the image, but to better associate images within semantic context surrounding it. In the case of the Darknet data, this context is often different proportions of selling, shipping, and item-quality terms.

The greater number of topics also creates many topics that cannot be deciphered based on the top terms and top images alone. Some single topics contain terms and images from a wide semantic range, including a mixture of pills, marijuana, counterfeit currency, and various software. These topics may represent a way to distinguish listings, but they do not provide much information on their own. It is possible that further cleaning of the text data could make these multimodal relationships more readily decipherable.

Retrieval Tasks One potential application of this type of neural network on Darknet marketplace datasets is the ability to see how different images relate in terms of predicted contextual text topics. Images of guns are near other images of guns, images of passports are near other images of passports, etc. This trained model allows not only for comparison of different images present within the dataset, but allows for querying of nearest neighbors based on any given image unseen during the training process. For example, Figure 11 illustrates that the 450-topic network predicts that an image of fire not in the training set would be seen in listings associated with images of pipes and marijuana. The network is capable of relating images in the context of the dataset, which could provide valuable threads to pull on and discover new insights.

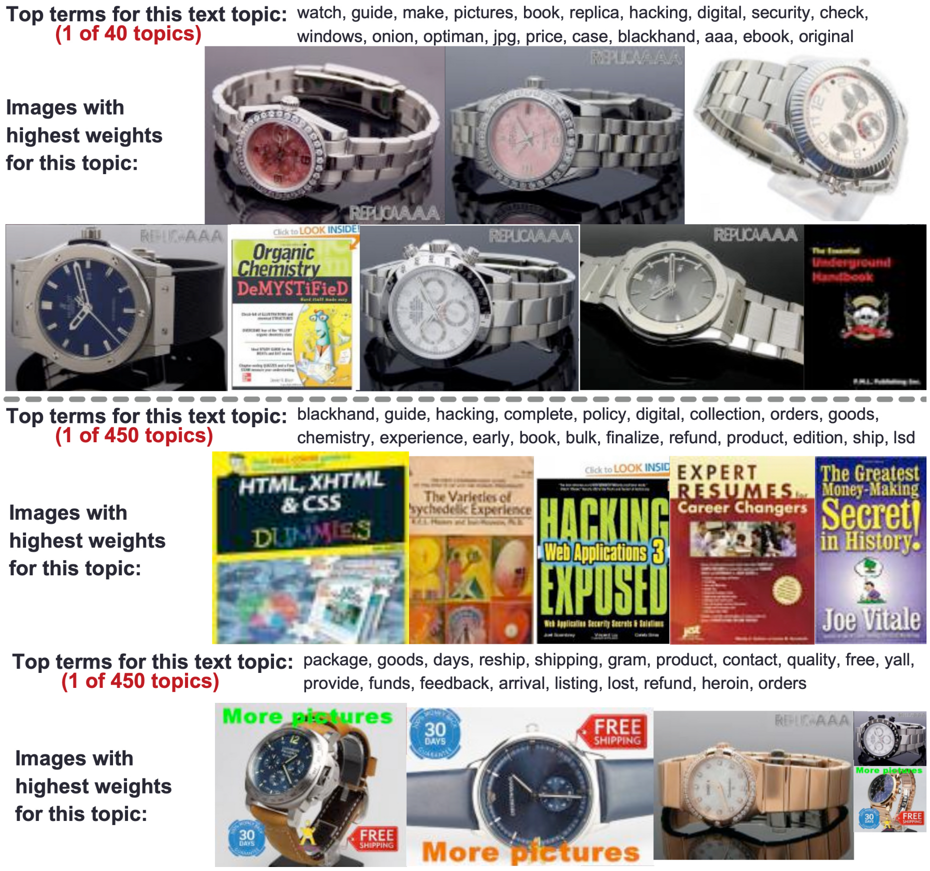

Top terms for this text topic:
(1 of 40 topics)
watch, guide, make, pictures, book, replica, hacking, digital, security, check, windows, onion, optiman, jpg, price, case, blackhand, aaa, ebook, original
Images with highest weights for this topic:
REPLICAAAA
Organic Chemistry DeMYSTiFieD
Click to LOOK INSIDE!
The Essential Underground Handbook
Top terms for this text topic:
(1 of 450 topics)
blackhand, guide, hacking, complete, policy, digital, collection, orders, goods, chemistry, experience, early, book, bulk, finalize, refund, product, edition, ship, lsd
Images with highest weights for this topic:
HTML, XHTML & CSS DUMMIES
The Varieties of Psychedelic Experience
HACKING Web Applications 3 EXPOSED
EXPERT RESUMES for Career Changers
The Greatest Money-Making Secret in History! Joe Vitale
Top terms for this text topic:
(1 of 450 topics)
package, goods, days, reship, shipping, gram, product, contact, quality, free, yall, provide, funds, feedback, arrival, listing, lost, refund, heroin, orders
Images with highest weights for this topic:
More pictures
30 DAYS
FREE SHIPPING
REPLICAAAA

Fig. 10. Larger-Dimension Topic Models lead to more specific terms and images. The 40-topic model had a topic for books and watches, while the 450-topic model had separate topics for watches and books individually.

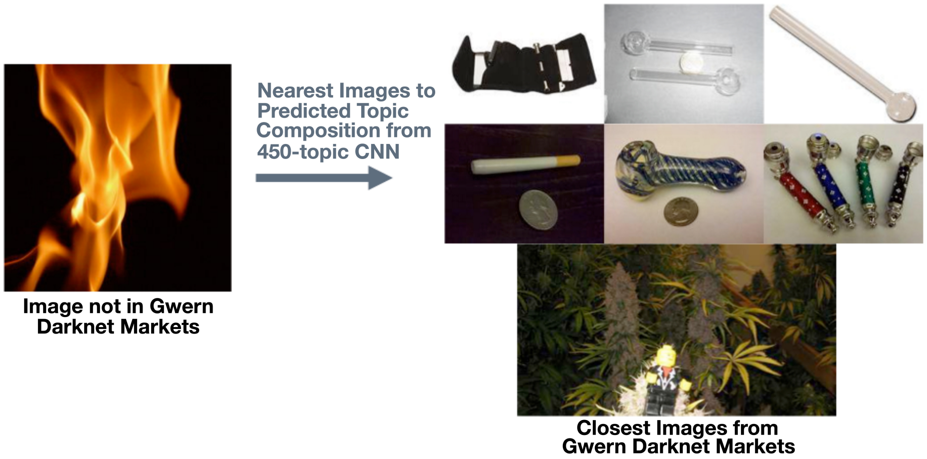
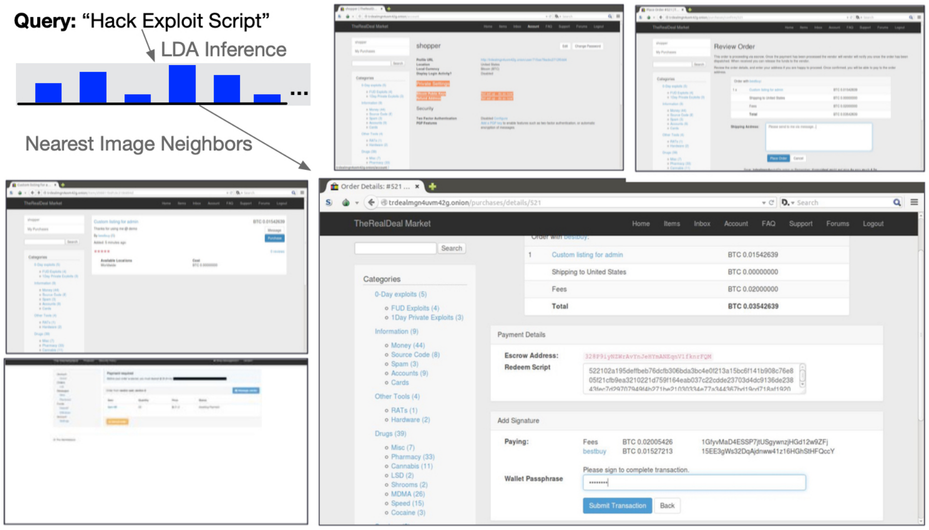

Nearest Images to Predicted Topic Composition from 450-topic CNN
Image not in Gwern Darknet Markets
Closest Images from Gwern Darknet Markets

Fig. 11. Image of Fire (not in the dataset) is predicted to be related to Images from the Gwern Darknet Markets Archive that would be in Listings of the same Textual Topic Composition (i.e. Fire and Smoking)



Order with bestbuy:

1	Custom listing for admin	BTC 0.01542639
	Shipping to United States	BTC 0.00000000
	Fees	BTC 0.02000000
	Total	**BTC 0.03542639**

Payment Details

Escrow Address: 328P9iyNZWrAvYnJeHYmANEqnV1fknrFQM

Redeem Script 522102a195deffbeb76dcfb306bda3bc4e0f213a15bc6f141b908c76e8 05f21cfb9ea3210221d759f164eab037c22cdde23703d4dc9136de238

Add Signature

Paying:	Fees	BTC 0.02005426	1GfyvMaD4ESSP7jtUSgywnzjHGd12w9ZFj
	bestbuy	BTC 0.01527213	15EE3gWs32DqAjdnww41z16HGhStHFQccY

Please sign to complete transaction.

Wallet Passphrase ••••••••

Fig. 12. The LDA-inferred topic composition of the query "hack exploit script" is nearest to predicted topic-compositions (450 dimensions) of screenshots, originating from tutorials detailing steps one would take to purchase a script

These threads can be found by utilizing LDA to infer topic distributions based on arbitrary text, and finding nearest neighbors to that text. If an analyst is curious what images would be associated with any collection of terms in the dataset, they can utilize the LDA model to infer the topic distribution for those terms and then find the nearest neighboring images in that space via Jensen-Shannon Divergence [14] the similarity measure. For example, as shown in Figure 12, an analyst curious about finding hacker materials on the darknet could query "hack exploit script", which retrieves nearest neighbor images that are screenshots that are found within listings. In this case, it is not immediately obvious why the text and the images are related. Upon further analysis of the source listing the screenshots originated from, we see that these screenshot images are part of tutorials explaining how a customer can purchase scripts. From this observation, we can infer that listings that sell scripts and have screenshots may be catering to less experienced customers who may not know how to exchange cryptocurrency for illicit products. Automated predicting of images' positions in the LDA textual topic space allows for analysts to create and refine queries that can lead to revealed relations and further insights into ever-changing listing categories on DNMs.

5 Discussion and Future Work

The technique of creating image descriptors to associate with textual semantics presents ways to summarize large multimodal datasets without much initial human input. Adding images to visualizations of textual topics could help analysts obtain a more-complete view of what a dataset contains. In addition, the ability to query based on images may allow analysts to help find items that may be more difficult to describe with words accurately.

The methods demonstrated in this paper are not for finding a needle in the haystack, but for finding if a certain color of hay is mixed within the haystack. Viewing topic summaries, and querying all of the topic-inferred media, does not present analysts with complete knowledge of datasets' outliers. The goal is to further support analysts ability to pull at strings, further learning about the composition until they can form their own generalizations about the data and deepen their explorations. More tools generated from the CNN presented in this paper could help further analysis once analyst generalizations are settled. Image retrieval based on image similarity, not necessarily semantic similarity could be done by retrieving based on the top pooling layer output of the CNN before the fully-connected layers [7, 19]. Classifiers could be built with Support Vector Machines or newly trained fully-connected layers, training on the pooling layer output, which is fine-tuned to distinguish visual features within the dataset. These classifiers could identify hand-labeled instances and future data of interest, such as classify images from the TOR Image Categories dataset [6]. Building supervised classifiers would likely take less hand-labeling time and less model training time than what is needed to train a neural network from scratch [7, 19]. To improve accuracy of these classifiers, it may help to retrain the neural network without transfer-learning as the network weights may converge to a saddle point based on the visual features of the pre-trained classes. This is because the CNN in this paper was pre-trained with weights to classify ImageNet [5] and thus would be expected to distinguish images in the new datasets via visual features useful for distinguishing ImageNet classes. To distinguish images that are not similar to ImageNet classes, it may be useful to transfer-learn from a network trained to distinguish similar classes, or train the network from a less biased initial state. Examples of domains this strategy may be useful within include but are not limited to: classifying handwriting styles from text written, classifying more abstract illustrations based on descriptive text, and classifying faces from biographic text. That being said, when classifying everyday items, ImageNet weights will likely perform well. The methodology demonstrated in this paper can be extended for supervised classification, while simultaneously allowing for visualizations of domain-specific (e.g. DNM listings) text-image relations to aid analysts in identifying existing and emerging semantic areas of interest.

Many datasets, especially collections of web resources, are multimodal and lend themselves to similar methods of self-supervised learning as mentioned in this paper. Future work could also investigate how other media, and other neural network architecture, could relate more media forms to LDA topic space. This method may be able to bootstrap analysis and classification of many domains and media, potentially linking text, images, video, and audio all to the same semantic space for a more complete view of any given dataset.

Future work may identify how to adapt existing LDA visualization strategies [22, 20, 21] to include images in conjunction with the existing textual elements. This could allow for analysts to not only better understand a dataset from LDA-categorization of text and images, but allow for greater exploration of hypothesis' based on the multimodal nature of many datasets. To better understand web content such as Darknet Markets, analysts will greatly benefit from tools that can automatically summarize and can enable the exploration of inter-relations between modalities of a dataset. In this work we propose initial tools for relatively computationally inexpensive, automatic semantic classification of multimodal datasets, and demonstrate these tools' abilities via Darknet Market data.

6 Conclusion

This paper presents tools to support analysts in quickly associating images with textual semantic features, within the context of a multimodal dataset. As demonstrated on Wikipedia and Darknet Market data, one can obtain relatively-fast automatic semantic classification of text and images, with almost no human prepocessing or labeling of the data. This can give analysts more complete knowledge of topics that compose a dataset, where text alone is not always fully descriptive of the multimodal documents' contents. This is especially useful in cases where the text is not very informative on its own, such as the Darknet Market listings analyzed in this paper, where many terms are artifacts of documents that can not be trivially removed. Beyond summarization, we demonstrate how the the trained CNN and LDA models in conjunction can retrieve documents and images based on both text and image queries. This can allow analysts to better explore areas not obvious in the topic model, such as the relation between tutorials and hacking scripts found in the Darknet. Additionally, almost no domain knowledge is needed to train these models, meaning the methods in this paper could serve as a "launching point" into exploration of unorganized multimodal datasets. Future work could extend tools and workflows taken from this launching point into more fine-grain tools for supervised classification and visualization of multimodal relations of the data. The combination of transfer-learning and unsupervised topic modeling result in a relatively-quick method for analysts to gain initial understanding of large mutimodal datasets' compositions, such as Darknet Markets, without requiring significant analyst experience in the domain of the dataset.

References

1. How the feds took down the Dread Pirate Roberts — Ars Technica, https://arstechnica.com/tech-policy/2013/10/how-the-feds-took-down-the-dread-pirate-roberts/
2. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. jmlr.org http://www.jmlr.org/papers/v3/blei03a.html
3. Buhrmester, M., Kwang, T.: Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? journals.sagepub.com http://journals.sagepub.com/doi/abs/10.1177/1745691610393980
4. Chollet, F.: Deep learning with python (2017), https://dl.acm.org/citation.cfm?id=3203489
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. Tech. rep., http://www.image-net.org.
6. Fidalgo, E., Alegre, E., González-Castro, V., Fernández-Robles, L.: Illegal Activity Categorisation in DarkNet Based on Image Classification Using CREIC Method. pp. 600–609 (2018). https://doi.org/10.1007/978-3-319-67180-2_58
7. Gomez, L., Patel, Y., Rusiñol, M.: Self-supervised learning of visual features through embedding images into text topic spaces. openaccess.thecvf.com http://openaccess.thecvf.com/content_cvpr_2017/papers/Gomez_Self-Supervised_Learning_of_CVPR_2017_paper.pdf
8. Grant, S., Cordy, J.: Estimating the optimal number of latent concepts in source code analysis. ieeexplore.ieee.org https://ieeexplore.ieee.org/abstract/document/5601828/
9. Gwern Branwen, Nicolas Christin, David Décary-Hétu, Rasmus Munksgaard Andersen, StExo, El Presidente, Anonymous, Daryl Lau, Sohhlz, Delyan Kratunov, Vince Cakic, Van Buskirk, Whom, Michael McKenna, Sigi Goode: Dark Net Market archives (2015), https://www.gwern.net/DNM-archives
10. Gwern Branwen, Nicolas Christin, D.D.H.R.M.A.S., El Presidente, Anonymous, D.L.S.D.K.V.C.V.B.W.M.M.S.G.: Dark Net Market archives, 2011-2015 (2015), https://www.gwern.net/DNM-archives
11. He, K., Zhang, X., Ren, S., Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. Tech. rep., https://arxiv.org/pdf/1502.01852v1.pdf

12. Investigation, K.P.D., 2018, u.: Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling. Elsevier https://www.sciencedirect.com/science/article/pii/S1742287618302020
13. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. papers.nips.cc http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks
14. Lin, J.: Divergence measures based on the Shannon entropy. ieeexplore.ieee.org https://ieeexplore.ieee.org/abstract/document/61115/
15. Matilla, D., González-Castro, V., Fernández-Robles, L., Fidalgo, E., Al-Nabki, W.: Color SIFT Descriptors to Categorize Illegal Activities in Images of Onion Domains. Tech. rep., https://www.torproject.org/
16. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit (2002), http://mallet.cs.umass.edu
17. Munksgaard, R., Demant, J., Branwen, G.: Commentary A replication and methodological critique of the study "Evaluating drug trafficking on the Tor Network". International Journal of Drug Policy **35**, 92–96 (2016). https://doi.org/10.1016/j.drugpo.2016.02.027, http://dx.doi.org/10.1016/j.drugpo.2016.02.027
18. Pan, S.J., Yang, Q.: A Survey on Transfer Learning (2009). https://doi.org/10.1109/TKDE.2009.191, http://socrates.acadiau.ca/courses/comp/dsilver/NIPS95
19. Patel, Y., Gomez, L., Gomez, R., Rusiñol, M., Karatzas, D., Jawahar, C.V.: TextTopicNet - Self-Supervised Learning of Visual Features Through Embedding Images on Semantic Text Spaces (7 2018), http://arxiv.org/abs/1807.02110
20. Paul, C.L., Chang, J., Endert, A., Cramer, N., Gillen, D., Hampton, S., Burtner, R., Perko, R., Cook, K.A.: TexTonic: Interactive visualization for exploration and discovery of very large text collections. Information Visualization p. 147387161878539 (7 2018). https://doi.org/10.1177/1473871618785390, http://journals.sagepub.com/doi/10.1177/1473871618785390
21. Sievert, C., Shirley, K.E.: LDAvis: A method for visualizing and interpreting topics. Tech. rep. (2014), http://www.aclweb.org/anthology/W14-3110
22. Singh, J., Zerr, S., Siersdorfer, S.: Structure-Aware Visualization of Text Corpora. In: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval - CHIIR '17. pp. 107–116. ACM Press, New York, New York, USA (2017). https://doi.org/10.1145/3020165.3020182, http://dl.acm.org/citation.cfm?doid=3020165.3020182
23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception Architecture for Computer Vision (12 2015), http://arxiv.org/abs/1512.00567
24. Tsikrika, T., Popescu, A., Kludas, J.: Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011. ims-sites.dei.unipd.it http://ims-sites.dei.unipd.it/documents/71612/86377/CLEF2011wn-ImageCLEF-TsikrikaEt2011.pdf
25. Yosinski, J., Clune, J., Bengio, Y.: How transferable are features in deep neural networks? papers.nips.cc http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-n%E2%80%A6