The Language of Legal and Illegal Activity on the Darknet

Leshem Choshen, Dan Eldad, Daniel Hershcovich, Elior Sulem and Omri Abend

ACL 2019

האוניברסיטה העברית בירושלים THE HEBREW UNIVERSITY OF JERUSALEM



What is the Dark Web?

Public Web

Information that you would normally find on search engines.

Deep Web

Information that is not indexed by search engines and does not require authentication.

Dark Web

Information that is not accessible by normal internet browsers.

Daniel Hershcovich

2/28

Darknet

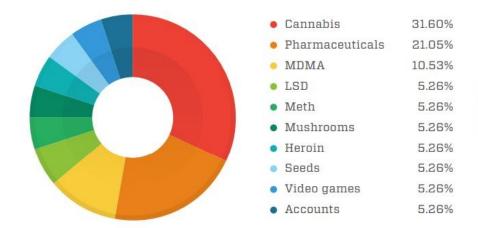
Used interchangeably in this work:

- Dark Web
- Darknet
- Tor network (Tor: an encrypted browser)
- Onion network (.onion top-level domain)



Hosts: onion services (hidden services).

Darknet Markets



¹Paganini (2015). "The Deep Web and Its Darknets". <□ > <♂ > < ≥ > <≥ > > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥ > < ≥

Language of the Darknet

How well do NIP tools work on Darknet text?

















Language of the Darknet

How well do NIP tools work on Darknet text?

















Can we automatically identify illegal activity?



Outline

Data: Darknet & eBay

2 Domain Differences: Vocabulary & Named Entities

Classification: Legal & Illegal Drugs, eBay

4 Cross-Domain Classification: Legal & Illegal Forums

DUTA-10K

Dataset of 10367 Onion Services text pages [Al Nabki et al., 2019].

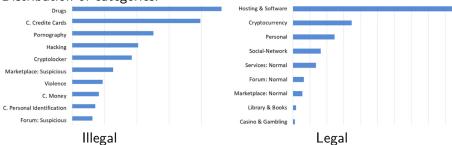
- 20% categorized as illegal and 48% as legal (32% unavailable).
- Of the illegal (suspicious) websites, 23% concern illegal drugs.

DUTA-10K

Dataset of 10367 Onion Services text pages [Al Nabki et al., 2019].

- 20% categorized as illegal and 48% as legal (32% unavailable).
- Of the illegal (suspicious) websites, 23% concern illegal drugs.

Distribution of categories:



4□ > 4□ > 4 = > 4 = > = 90

Control Data: eBay

Product descriptions acquired by searching for drug-related terms. Do not sell actual drugs, but rather drug-related products.



3 Layers Chip Style Herb Herbal Tobacco Grinder Weed Grinders

Description: Quantity: 1

Type : Tobacco Crusher

Feature: Stocked, Eco-Friendly

Material: plastic Size: 42*26mm Package include:

1PC Tobacco Crusher

Data

| | Public Web | Dark Web | | | |
|---------|---------------------------|------------------------------|--|--|--|
| Legal | еВау | Legal Onion | | | |
| Legai | (188 pages, 35,799 words) | (35 pages, 61,655 words) | | | |
| | | | | | |
| Illegal | | Illegal Onion | | | |
| megai | | (255 pages, 1,438,351 words) | | | |

Cleaning

- Remove non-linguistic content: buttons, encryption keys, URLs...
- Split to paragraphs, join to single lines, remove duplicates.

Finest organic cannabis grown by proffessional growers in the netherlands.

We double seal all packages for odor less delivery.

Shipping within 24 hours!

| Shipping within 24 hours! | | |
|---------------------------|----------------------|-------------|
| Product | Price | Quantity |
| 1g Original Haze | 15 EUR = 0.025 \$ 1_ | X Buy now |
| 5g Original Haze | 65 EUR = 0.108 B 1_ | X Buy now |
| 1g Bubblegum | 10 EUR = 0.017 B 1_ | X Buy now |
| 5g Bubblegum | 45 EUR = 0.075 \$ 1_ | X Buy now |
| 1g Jack Herer | 14 EUR = 0.023 \$ 1_ | X Buy now |
| 5g Jack Herer | 60 EUR = 0.099 \ 1_ | X Buy now |
| 1g Chronic | 9 EUR = 0.015 B 1_ | X Buy now |
| 5g Chronic | 40 EUR = 0.066 \$ 1_ | X Buy now |
| 1g Banana Kush | 11 EUR = 0.018 B 1_ | X Buy now |
| | 4 🗆 🕽 4 🗇 🖒 4 🖹 🔊 | 4 = b = 000 |

Clean Data

Sampled 571 paragraphs from each, for comparable size.

| | Public Web | Dark Web | |
|---------|----------------|----------------|--|
| Legal | eBay | Legal Onion | |
| | (14,276 words) | (14,802 words) | |
| | | | |
| Illegal | | Illegal Onion | |
| cgai | | (15,049 words) | |

Outline

- Data: Darknet & eBay
- 2 Domain Differences: Vocabulary & Named Entities

3 Classification: Legal & Illegal Drugs, eBay

4 Cross-Domain Classification: Legal & Illegal Forums

Vocabulary

Distance between word frequencies distributions, measured by:

Jensen-Shannon divergence

L1 distance

Splitting each dataset in half, we find small "self-distances"

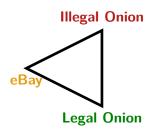
Vocabulary

Distance between word frequencies distributions, measured by:

Jensen-Shannon divergence

L1 distance

Splitting each dataset in half, we find small "self-distances", but the different domains are about equidistant.



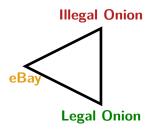
Vocabulary

Distance between word frequencies distributions, measured by:

Jensen-Shannon divergence

L1 distance

Splitting each dataset in half, we find small "self-distances", but the different domains are about equidistant.



Legal and illegal Onion should be considered different domains.

A D P A DP A E P A E P A E P A E P A E P A E P A E P A E P A E P A E P A E P A E P A E P A E P A E P A E P A E

Characteristics of Darknet Data

Diverse: sub-domains are distinguishable.

Unique: distinguishable from other domains.



Named Entities and Wikification

NE extraction [spaCy] + Wikification [Bunescu and Pașca, 2006].

| | % (of detected NEs) Wikifiable |
|----------------------|--------------------------------|
| eBay | 38.6 ± 2.00 |
| Illegal Onion | 32.5 ± 1.35 |
| Legal Onion | 50.8 ± 2.31 |

By manual inspection, NE precision and recall are low for **Illegal Onion**. For example: slang words for drugs (e.g., "kush") falsely picked up as NEs.

⇒ Standard NLP is not suited for this domain.

Daniel Hershcovich 15 / 28

Outline

- Data: Darknet & eBay
- 2 Domain Differences: Vocabulary & Named Entities

3 Classification: Legal & Illegal Drugs, eBay

4 Cross-Domain Classification: Legal & Illegal Forums

Classes

We identified three domains. Two binary classification settings:

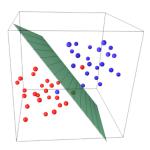
```
{ eBay, Legal Onion }
{ Legal Onion, Illegal Onion }
```

Classes

We identified three domains. Two binary classification settings:

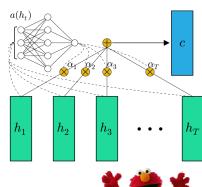
```
{ eBay, Legal Onion }
{ Legal Onion, Illegal Onion }
```

What are the linguistic features distinguishing them?



Classifiers

- NB: Naive Bayes (bag of words)
- SVM: Support Vector Machine
- BoE: sum/average GloVe + MLP
- seq2vec: BiLSTM + MLP
- attention: ELMo + BCN (self-attention)





Manipulations¹

- Full original text
- Drop content words
- Replace content words with their POS
- Drop **function** words
- Replace function words with their POS

{ADJ, ADV, NOUN, PROPN, VERB, X, NUM}

Manipulations

- Full original text
- Drop content words
 Replace content words with their POS
- Drop function words
 Replace function words with their POS

```
{ADJ, ADV, NOUN, PROPN, VERB, X, NUM}
```

```
Generic Viagra (Oral Jelly) is used for Erectile Dysfunction PROPN PROPN (PROPN PROPN) VERB VERB for PROPN PROPN
```

Manipulations

- Full original text
- Drop content words
 Replace content words with their POS
- Drop function words
 Replace function words with their POS

```
\{ADJ, ADV, NOUN, PROPN, VERB, X, NUM\}
```

```
Generic Viagra (Oral Jelly) is used for Erectile Dysfunction PROPN PROPN (PROPN PROPN) VERB VERB for PROPN PROPN
```

```
Welcome to SnowKings Good Quality Cocaine!

VERB to PROPN PROPN PROPN!
```

Results: eBay vs. Legal Onion Drugs

Clear separation by content (by NB), but also by function (by SVM).

| | full | drop content | drop function | pos content | pos function |
|--------------------------|------|-----------------|------------------|----------------|-----------------|
| NB | 91.4 | 57.8 | 90.5 | 56.9 | 92.2 |
| SVM | 63.8 | 64.7 | 63.8 | 68.1 | 63.8 |
| BoE_{sum} | 66.4 | 56.0 | 63.8 | 50.9 | 76.7 |
| BoE_{average} | 75.0 | 55.2 | 59.5 | 50.0 | 75.0 |
| seq2vec | 73.3 | 53.8 | 65.5 | 65.5 | 75.0 |
| attention | 82.8 | 57.5 | 85.3 | 62.1 | 82.8 |

Results: Legal vs. Illegal Onion Drugs

Harder to distinguish by content, easier by POS distribution (SVM again).

| | full | drop content | drop function | pos content | pos function |
|--------------------------|------|-----------------|------------------|----------------|-----------------|
| NB | 77.6 | 53.4 | 87.9 | 51.7 | 77.6 |
| SVM | 63.8 | 66.4 | 63.8 | 70.7 | 63.8 |
| BoE_{sum} | 52.6 | 61.2 | 74.1 | 50.9 | 51.7 |
| BoE_{average} | 57.8 | 57.8 | 52.6 | 55.2 | 50.9 |
| seq2vec | 56.9 | 55.0 | 54.3 | 59.5 | 49.1 |
| attention | 64.7 | 51.4 | 62.9 | 55.2 | 69.0 |

Daniel Hershcovich 21 / 28

Classification Challenges

Simple classifiers (NB, SVM) work best.

- Small training data.
- Non-standard language.
- Understudied domain.

Outline

Data: Darknet & eBay

2 Domain Differences: Vocabulary & Named Entities

3 Classification: Legal & Illegal Drugs, eBay

4 Cross-Domain Classification: Legal & Illegal Forums

Darknet Forums

Can we generalize beyond drugs?

Darknet Forums

Can we generalize beyond drugs?

DUTA-10K also contain Legal Forums and Illegal Forums. Multi-topic and user-generated.



Results: Legal vs. Illegal Onion Forums

Harder for most classifiers, but SVM succeeds using content and function.

| | full | drop content | drop function | pos content | pos function |
|--------------------------|------|-----------------|------------------|----------------|-----------------|
| NB | 74.1 | 50.9 | 78.4 | 50.9 | 72.4 |
| SVM | 85.3 | 75.9 | 56.0 | 81.9 | 81.0 |
| BoE_{sum} | 25.9 | 32.8 | 21.6 | 36.2 | 35.3 |
| BoE_{average} | 40.5 | 42.2 | 31.9 | 48.3 | 53.4 |
| seq2vec | 50.0 | 48.9 | 50.9 | 28.4 | 51.7 |
| attention | 31.0 | 37.2 | 33.6 | 27.6 | 30.2 |

Daniel Hershcovich 25 / 28

Results: Trained on Drugs, Tested on Forums

Effective cross-domain generalization even with bag-of-words.

| | full | drop content | drop function | pos content | pos function |
|--------------------------|------|-----------------|------------------|----------------|-----------------|
| NB | 78.4 | 63.8 | 89.7 | 63.8 | 79.3 |
| SVM | 62.1 | 69.0 | 54.3 | 69.8 | 62.1 |
| BoE_{sum} | 45.7 | 50.9 | 49.1 | 50.9 | 50.0 |
| BoE_{average} | 49.1 | 51.7 | 51.7 | 52.6 | 58.6 |
| seq2vec | 51.7 | 61.1 | 51.7 | 54.3 | 57.8 |
| attention | 65.5 | 59.2 | 65.5 | 50.9 | 66.4 |

Daniel Hershcovich 26 / 28

Differences between legal and illegal Darknet sites:

- Vocabulary
- Shallow syntax (POS)
- Named entities

Differences between legal and illegal Darknet sites:

- Vocabulary
- Shallow syntax (POS)
- Named entities

Identified by:

- Word statistics: diverse and unique
- Wikification: works less well on illegal
- Predictive: simple classifiers work best

Differences between legal and illegal Darknet sites:

- Vocabulary
- Shallow syntax (POS)
- Named entities

Identified by:

- Word statistics: diverse and unique
- Wikification: works less well on illegal
- Predictive: simple classifiers work best

Code: https://github.com/huji-nlp/cyber

Data: dan.eldad1@mail.huji.ac.il

Differences between legal and illegal Darknet sites:

- Vocabulary
- Shallow syntax (POS)
- Named entities

Identified by:

- Word statistics: diverse and unique
- Wikification: works less well on illegal
- Predictive: simple classifiers work best

Code: https://github.com/huji-nlp/cyber

Data: dan.eldad1@mail.huji.ac.il



Daniel Hershcovich 27 / 28

References I



Al Nabki, M. W., Fidalgo, E., Alegre, E., and Fernández-Robles, L. (2019).

ToRank: Identifying the most influential suspicious domains in the Tor network. Expert Systems with Applications, 123:212–226.



Bunescu, R. and Pașca, M. (2006).

Using encyclopedic knowledge for named entity disambiguation. In ${\it Proc. of EACL}$.