

ACL 2023

DarkBERT : A Language Model for the Dark Side of the Internet

Youngjin Jin, Eugene Jang, Jian Cui, Jin-Woo Chung, Yongjae Lee, Seungwon Shin

Presenter : Seungho Song

Introduction.

Surface Web

- Available to the general public using standard search engines.
- Can be accessed using standard web browsers that do not require any special configuration.

Deep Web

- Not indexed or searchable by ordinary search engines.
- They do not use common top-level domains (TLDs).

Dark Web

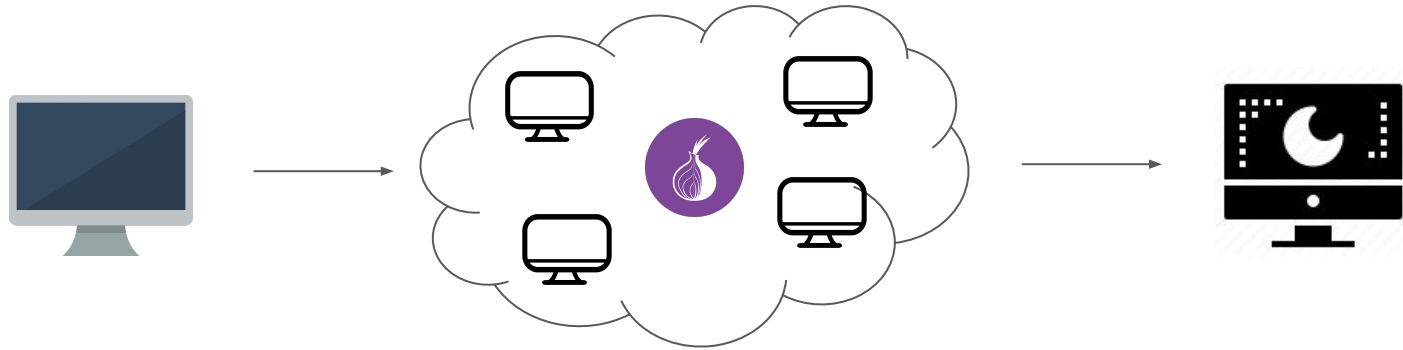
- Relies on connections made between trusted peers.
- Require specialized software, tools to access.



Introduction (cont.)

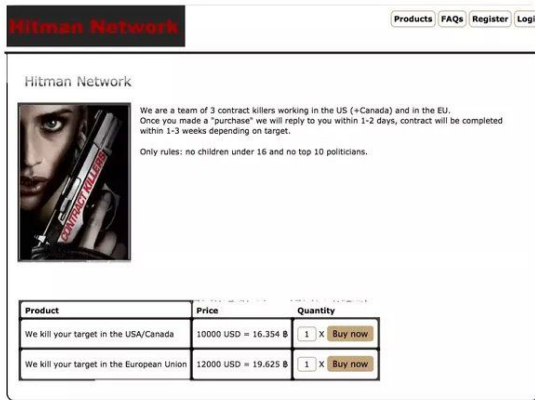
Dark Web

- Not Indexing. **Only assessed by specific tools** such as TOR, I2P, etc.



Introduction (cont.)

Dark websites example



Hitman Network

Products | FAQs | Register | Login

Hitman Network

We are a team of 3 contract killers working in the US (+Canada) and in the EU. Once you made a "purchase" we will reply to you within 1-2 days, contract will be completed within 1-3 weeks depending on target.

Only rules: no children under 16 and no top 10 politicians.

Product	Price	Quantity
We kill your target in the USA/Canada	10000 USD = 16,354 \$	1 x Buy now
We kill your target in the European Union	12000 USD = 19,625 \$	1 x Buy now

Contract killers



FakeID

Main | News | Services | Samples | faq | Order | Contacts

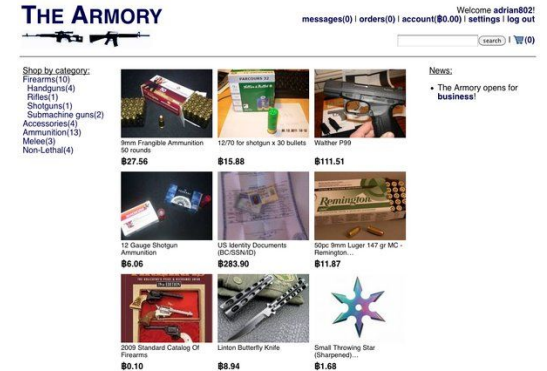
Welcome

Welcome to **Fake Documents Service**- the unique producer of quality fake documents. We offer only original high-quality fake passports, driver's licenses, ID cards, stamps and other products for following countries: Australia, Belgium, Brazil, Canada, Finland, France, Germany, Italy, Netherlands, UK, USA and some others.

If you want to learn more about what kinds of documents can be found in our website please visit the sections "Services" and "Samples". You can find more details about ordering procedure and additional details visiting the sections "FAQ" and "Order".

© 2015 Copyright Fake Documents.
For entertainment only. Not a government document.
Terms and Conditions

Fake Docs



THE ARMORY

messages(0) | orders(0) | account(\$0.00) | settings | log out

Shop by category:
Firearms(10)
Handguns(4)
Rifles(1)
Shotguns(1)
Submachine guns(2)
Accessories(4)
Ammunition(13)
Melee(3)
Non-Lethal(4)

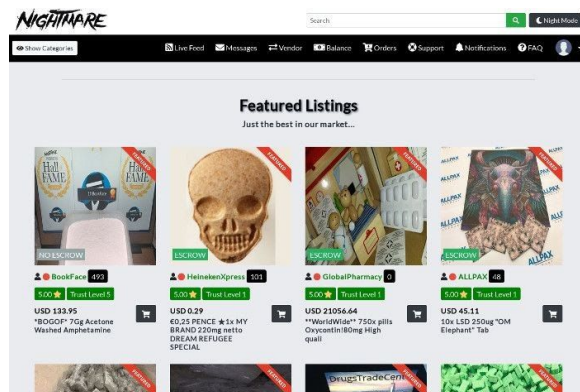
News:
• The Armory opens for business!

12 Gauge Shotgun Ammunition 50 rounds	12/70 for shotgun x 30 bullets	Walther P99
\$27.56	\$15.88	\$111.51
12 Gauge Shotgun Ammunition	US Identity Documents (20-identity)	50pc 9mm Luger 147 gr MC - Remington...
\$6.06	\$283.90	\$11.87
2009 Standard Catalog Of Firearms	Union Butterfly Knife	Small Throwing Star (Sharpened)...
\$0.10	\$9.94	\$1.68

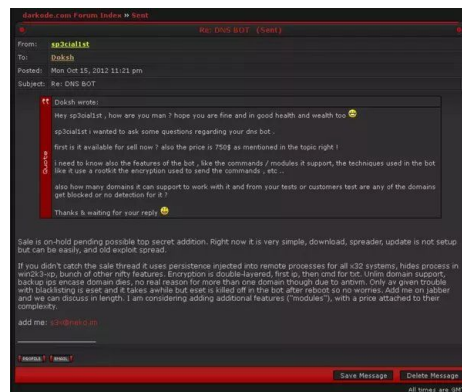
Weapons

Introduction (cont.)

Dark websites example (cont.)



Drugs



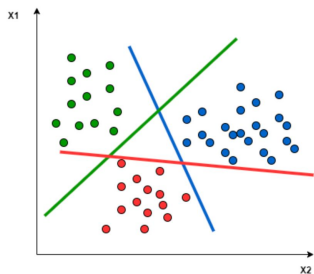
Malwares / Bot



Hacker Recruiting

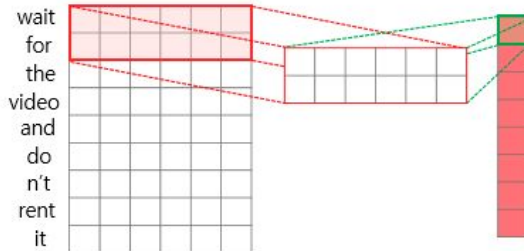
Introduction (cont.)

Solutions to protect malicious activities in Dark Web : NLP



- Multi-class SVM + BoW

- 단어 분포를 통한 비교
- 키워드 위주



- CNN + GloVe embedding

- GloVe : 단어당 의미를 벡터화
- CNN : 벡터끼리의 위치를 고려하는 모델
- 단어의 의미를 뽑아 문맥을 이해



- BERT

- 문맥을 통째로 벡터화하는 모델
- 문맥을 고려해 단어 의미 이해

Introduction (cont.)

BERT (Bidirectional Encoder Representations from Transformers)

- Masked Language Model

: Predict words that will appear in the middle of the current sentence.

ex. Paris is the [MASK] of France.

1. Initial version of Language models => banana
2. Updated version of Language models => president
- ...
3. **Enough updated** version of Language models => **capital**



Google BERT

Pretrained from **2.5B** English Wikipedia text + **0.8B** book text data

Introduction (cont.)

BERT (Bidirectional Encoder Representations from Transformers)

- Masked Language Model

: Predict words that will appear in the middle of the current sentence.

ex. Paris is the [MASK] of France.

1. Initial version of Language models => banana
2. Updated version of Language models => president
- ...
3. **Enough updated** version of Language models => **capital**



Google BERT

Pretrained from **2.5B** English Wikipedia text + **0.8B** book text data

→ **Surface Web Data**

Introduction (cont.)

Different word meaning between Surface vs Dark Web

ex.

- Weed => 잡초
- Tesla, Toyota => 자동차 브랜드
- wasabi => 고추냉이

Introduction (cont.)

Different word meaning between Surface vs Dark Web

ex.

- Weed => 잡초 | 대마초
- Tesla, Toyota => 자동차 브랜드 | 마약
- wasabi => 고추냉이 | **blockchain wallet**

Popular pre-trained language models such as **BERT** are not ideal for **predicting CTI contents** such as Dark Web browsing.

Introduction (cont.)

Different word meaning between Surface vs Dark Web

ex.

- Weed => 잡초 | 대마초
- Tesla, Toyota => 자동차 브랜드 | 마약
- wasabi => 고추냉이 | blockchain wallet

Popular pre-trained language models such as **BERT** are not ideal for predicting CTI contents such as Dark Web browsing.

Specific Domain



Need **domain-specific pretrained language models** that are **suitable for Dark Web domain**

Introduction (cont.)

Different word meaning between Surface vs Dark Web

ex.

- Weed => 잡초 | 대마초
- Tesla, Toyota => 자동차 브랜드 | 마약
- wasabi => 고추냉이 | blockchain wallet

Popular pre-trained language models such as **BERT** are not ideal for predicting CTI contents such as Dark Web browsing.

Specific Domain



Need **domain-specific pretrained language models** that are **suitable for Dark Web domain**
=> **DarkBERT**

DarkBERT

A new language model **pretrained on a Dark Web** corpus.

Capable of **representing the language used in Dark Web** domain
compared to that of the Surface Web



DarkBERT Construction

- Architecture Overview

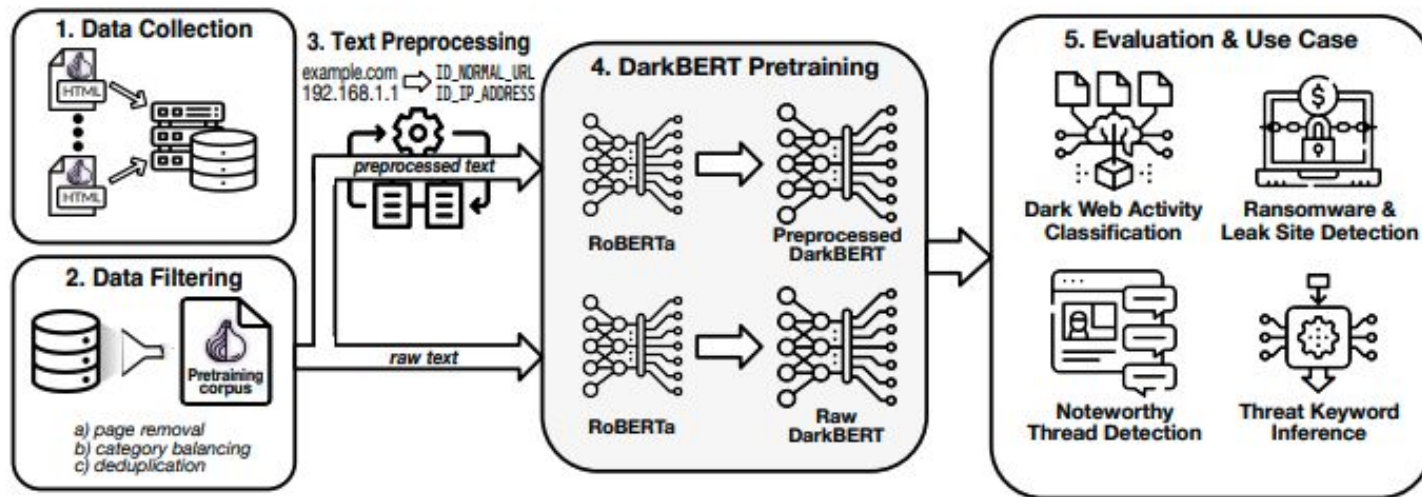


Figure 1: Illustration of the DarkBERT pretraining process and the various use case scenarios for evaluation.

DarkBERT Construction (cont)

1. Data Collection

- Method

- Collect seed addresses from Ahmia & public onion url lists
- Crawl the Dark Web datas & Parse into Title, Body
- Select only contents written in English

Table 8: Dark Web data collection statistics

Statistics	Value
Total number of collected pages	6.1 M
Average number of characters per page	7,980
Minimum number of characters in a page	7
Maximum number of characters in a page	17,786,986
Per-page character count statistics	Character count
Q_1 (25th quartile)	1,318
Q_2 (50th quartile)	2,581
Q_3 (75th quartile)	5,753

DarkBERT Construction (cont)

2. Data Filtering

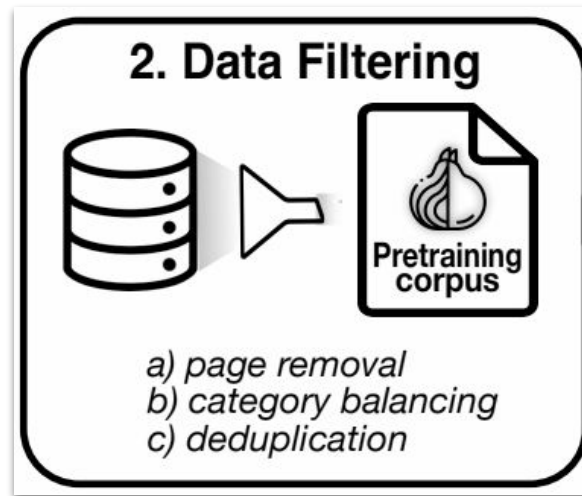
- **Purpose**

- remove unnecessary information

ex) error messages / blank pages / duplicates of other pages.

- **3 Key measures**

1. Removal of Pages with low information density
2. Category Balancing
3. Deduplication



DarkBERT Construction (cont)

2. Data Filtering

- 3 Key measures

1. **Removal of Pages with low information density**

- a. MIN : 500 char / MAX : 10,000 char
- b. Texts under MIN
 - i. Error messages (404 not found / please sign-in / Captcha Error ...)
- c. Texts over MAX
 - i. Unnecessary pages (continuous repetitions of certain strings)

DarkBERT Construction (cont)

2. Data Filtering

- 3 Key measures

2. Category Balancing

- Half of the pages on the dark web are **pornography** => PornBERT? 😱
- Remove several pornography datas using Category Classifier ([CoDA](#), 2022 ACL, Jin et al.)

Table 9: Dark Web page classification and pretraining data statistics. The statistics marked as *(full)* represent the original data collection, and *(pretraining)* represents the data after deduplication and category balancing are applied.

Category	Page Count (full)	Total Size (full)	Average Size per Page (full)	Page Count (pretraining)	Total Size (pretraining)	Deduplication Rate	Total Reduction Rate
Pornography	2,267,628	9.70 GB	4.28 KB	224,781	971.0 MB	2.91%	89.98%
Drugs	503,433	1.75 GB	3.47 KB	228,965	766.7 MB	23.31%	56.19%
Financial	637,917	2.10 GB	3.29 KB	253,171	874.1 MB	12.45%	58.38%
Gambling	43,041	0.15 GB	3.38 KB	40,584	137.5 MB	5.37%	5.37%
Cryptocurrency	412,349	1.36 GB	3.29 KB	249,811	897.6 MB	10.28%	34.00%
Hacking	801,330	3.51 GB	4.38 KB	57,183	242.7 MB	75.73%	93.09%
Arms / Weapons	46,616	0.14 GB	2.70 KB	43,250	129.9 MB	6.15%	6.15%
Violence	323,738	1.21 GB	3.74 KB	253,566	959.8 MB	4.02%	20.68%
Electronics	401,196	0.89 GB	2.21 KB	381,218	850.4 MB	4.17%	4.45%
Total	5,437,248	20.79 GB	-	1,732,529	5.83 GB	18.69%	71.96%

DarkBERT Construction (cont)

2. Data Filtering

- 3 Key measures

3. Deduplication

- Remove pages which are in duplicated category from others.

Table 9: Dark Web page classification and pretraining data statistics. The statistics marked as *(full)* represent the original data collection, and *(pretraining)* represents the data after deduplication and category balancing are applied.

Category	Page Count (full)	Total Size (full)	Average Size per Page (full)	Page Count (pretraining)	Total Size (pretraining)	Deduplication Rate	Total Reduction Rate
Pornography	2,267,628	9.70 GB	4.28 KB	224,781	971.0 MB	2.91%	89.98%
Drugs	503,433	1.75 GB	3.47 KB	228,965	766.7 MB	23.31%	56.19%
Financial	637,917	2.10 GB	3.29 KB	253,171	874.1 MB	12.45%	58.38%
Gambling	43,041	0.15 GB	3.38 KB	40,584	137.5 MB	5.37%	5.37%
Cryptocurrency	412,349	1.36 GB	3.29 KB	249,811	897.6 MB	10.28%	34.00%
Hacking	801,330	3.51 GB	4.38 KB	57,183	242.7 MB	75.73%	93.09%
Arms / Weapons	46,616	0.14 GB	2.70 KB	43,250	129.9 MB	6.15%	6.15%
Violence	323,738	1.21 GB	3.74 KB	253,566	959.8 MB	4.02%	20.68%
Electronics	401,196	0.89 GB	2.21 KB	381,218	850.4 MB	4.17%	4.45%
Total	5,437,248	20.79 GB	-	1,732,529	5.83 GB	18.69%	71.96%

DarkBERT Construction (cont)

3. Text Pre-processing

- Purpose

- To address **ethical** considerations.

- : Most model doesn't learn representations from sensitive information.

- Methods

1. Identifier Masking

2. Removing entire text

Table 10: The types of identifier masks and the list of preprocessed texts.

Identifier Type	Example Text or Description	Preprocess Action Type	Identifier Mask Token
Email Addresses	example@email.com	Replace with token	ID_EMAIL
URLs (non-onion domain)	www.example.com	Replace with token	ID_NORMAL_URL
URLs (onion domain)	<u>https://www.example.com/home</u> <u>facebookwkhpilnemxj7asaniu7vnjjbiltxjqhye3mhbshg7kx5tfyd.onion</u>	Replace with token	ID_ONION_URL
IP Addresses (IPv4 & IPv6)	192.168.1.1 fe80::1ff:fe23:4567:890a%eth2	Replace with token	ID_IP_ADDRESS
Cryptocurrency Addresses	BTC, ETH, LTC addresses	Replace with token	ID_BTC_ADDRESS ID_ETH_ADDRESS ID_LTC_ADDRESS
Lengthy "Words"	Any group of non-whitespace characters that are 38 or more letters long	Replace with token	ID_LONGWORD
Uncommon Characters	Any characters out of Unicode range from U+0000 to U+00FF	Remove from text	-
Whitespaces	Newline characters, tabs, spaces, etc.	Truncate to a single space	-

DarkBERT Construction (cont)

4. Pre-training

- Two versions of DarkBERT : **raw (masking x) + preprocessed (masking o)**

Table 1: The two variations of Dark Web text corpus used to train DarkBERT.

Corpus	Data Size	Time Taken to Pretrain DarkBERT
Raw Text	5.83 GB	367.4 hours (15.31 days)
Preprocessed Text	5.20 GB	361.6 hours (15.07 days)

- Base model
 - **RoBERTa** : improved version of BERT (made by Meta)
 - More Data, More learning time
 - 6,100,000 pages training (15 days learning)

DarkBERT Construction (cont)

4. Pre-training

- **Tokenization**
 - Use RoBERTa's same BPE (Byte-Pair Encoding) tokenization vocabulary
- Every settings without preprocessing between 2 versions of DarkBERT are **SAME**.

Table 11: The hyperparameters used for pretraining the two versions of DarkBERT.

Hyperparameter	Value
Number of Layers	12
Hidden Size	768
Feedforward NN	
Inner Hidden Size	3072
Attention Heads	12
Attention Head Size	64
Dropout	0.1
Attention Dropout	0.1
Max Sequence Length	512
Warmup Steps	24000
Peak Learning Rate	6e-4
Batch Size	8192
Weight Decay	0.01
Max Steps	20K
Learning Rate Decay	Linear
Adam ϵ	1e-6
Adam β_1	0.9
Adam β_2	0.98
Gradient Clipping	0.0

Evaluation

Benchmark

- Experiment : **Dark Web activity Classification**
- Models : BERT, RoBERTa

Datasets

- DUTA (**D**ark web **U**sage **T**ext **A**ddress)
 - => need to pre-processing
- CoDA (**C**omprehensive **D**ark web **A**nnotations)
 - => no need to pre-processing

Table 2: Dataset statistics used for Dark Web activity categorization.

DUTA (DUTA-10K)		CoDA	
Category	Page count	Category	Page count
Hosting & Software	1949	Others	2131
Cryptocurrency	798	Pornography	1171
Down	714	Drugs	967
Locked	682	Financial	956
Personal	419	Gambling	756
Counterfeit Credit Cards	392	Crypto	745
Social Network	293	Hacking	630
Drugs	290	Arms	597
Services	284	Violence	482
Pornography	226	Electronics	420
Marketplace	189		
Hacking	182		
Forum	128		
Total	6524	Total	8855

Evaluation

Experimentation

- **Experimental**
 - DarkBERT : raw version, preprocessed version
- **Control**
 - BERT : cased model, uncased model
 - RoBERTa : cased model, uncased model

Evaluation

Results

1. **DarkBERT outperforms** other language models (BERT, RoBERTa),
But **both BERT and RoBERTa exhibit relatively similar performances** to DarkBERT.

Table 3: Dark Web activity classification evaluation results. Boldface indicates best performance.

Dataset	Model	Precision	Recall	F1 score	Dataset	Model	Precision	Recall	F1 score
DUTA _{cased}	BERT _{cased}	77.31	76.91	77.09	CoDA _{cased}	BERT _{cased}	92.12	92.16	92.13
	BERT _{uncased}	78.21	78.20	78.19		BERT _{uncased}	92.83	92.67	92.75
	RoBERTa	78.54	78.79	78.63		RoBERTa	93.36	93.27	93.31
	DarkBERT _{raw}	80.11	79.94	80.01		DarkBERT _{raw}	94.15	94.35	94.25
	DarkBERT _{preproc}	79.90	80.08	79.98		DarkBERT _{preproc}	94.26	94.33	94.29
DUTA _{uncased}	BERT _{cased}	78.11	77.97	77.99	CoDA _{uncased}	BERT _{cased}	92.86	92.85	92.85
	BERT _{uncased}	78.21	78.20	78.19		BERT _{uncased}	92.83	92.67	92.75
	RoBERTa	78.42	78.36	78.37		RoBERTa	93.30	93.40	93.34
	DarkBERT _{raw}	79.47	79.49	79.47		DarkBERT _{raw}	94.46	94.45	94.46
	DarkBERT _{preproc}	79.65	79.77	79.69		DarkBERT _{preproc}	94.31	94.53	94.42

Evaluation

Results

2. All language models perform significantly **better for the CoDA**, compared to DUTA

Table 3: Dark Web activity classification evaluation results. Boldface indicates best performance.

Dataset	Model	Precision	Recall	F1 score	Dataset	Model	Precision	Recall	F1 score
DUTA _{cased}	BERT _{cased}	77.31	76.91	77.09	CoDA _{cased}	BERT _{cased}	92.12	92.16	92.13
	BERT _{uncased}	78.21	78.20	78.19		BERT _{uncased}	92.83	92.67	92.75
	RoBERTa	78.54	78.79	78.63		RoBERTa	93.36	93.27	93.31
	DarkBERT _{raw}	80.11	79.94	80.01		DarkBERT _{raw}	94.15	94.35	94.25
	DarkBERT _{preproc}	79.90	80.08	79.98		DarkBERT _{preproc}	94.26	94.33	94.29
DUTA _{uncased}	BERT _{cased}	78.11	77.97	77.99	CoDA _{uncased}	BERT _{cased}	92.86	92.85	92.85
	BERT _{uncased}	78.21	78.20	78.19		BERT _{uncased}	92.83	92.67	92.75
	RoBERTa	78.42	78.36	78.37		RoBERTa	93.30	93.40	93.34
	DarkBERT _{raw}	79.47	79.49	79.47		DarkBERT _{raw}	94.46	94.45	94.46
	DarkBERT _{preproc}	79.65	79.77	79.69		DarkBERT _{preproc}	94.31	94.53	94.42

Evaluation

Results

3. In performance of Misclassification, **DarkBERT show the best performance** than others.
But some categories show very similar performances across all four models.

Evaluation

Results

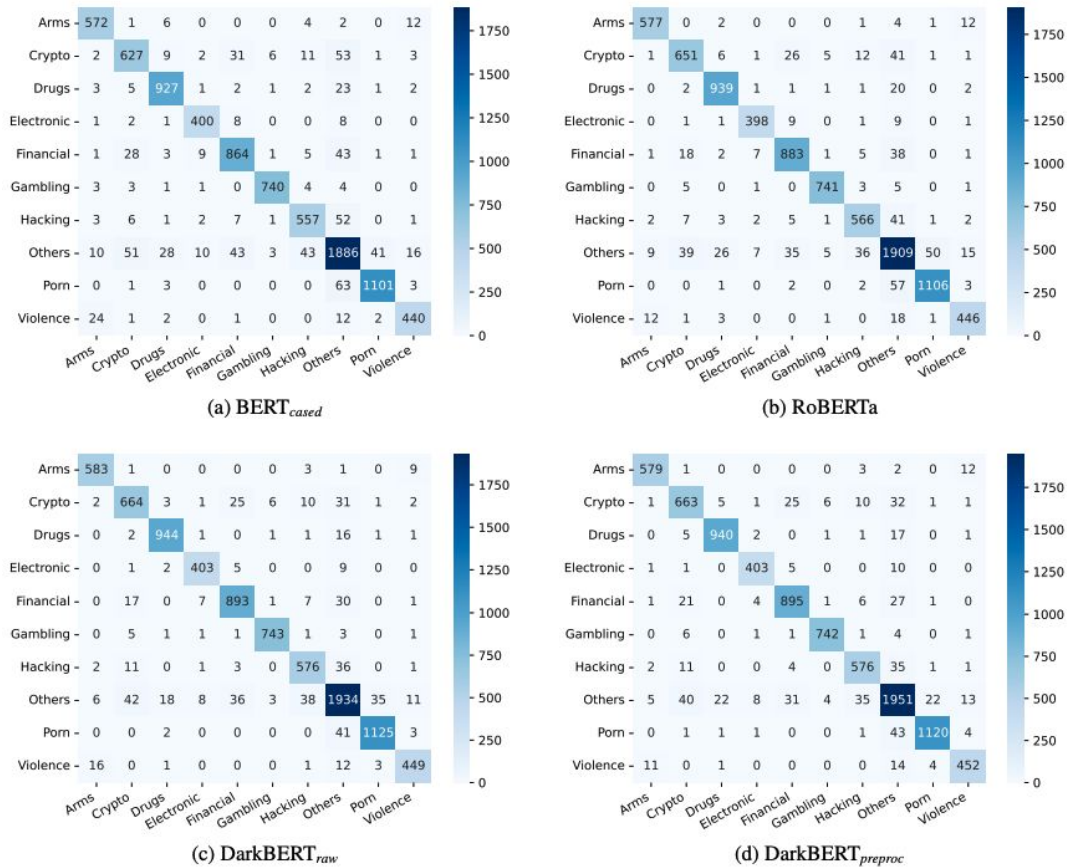


Figure 8: Confusion matrices for selected language models evaluated on the CoDA_{cased} dataset

Evaluation

Results

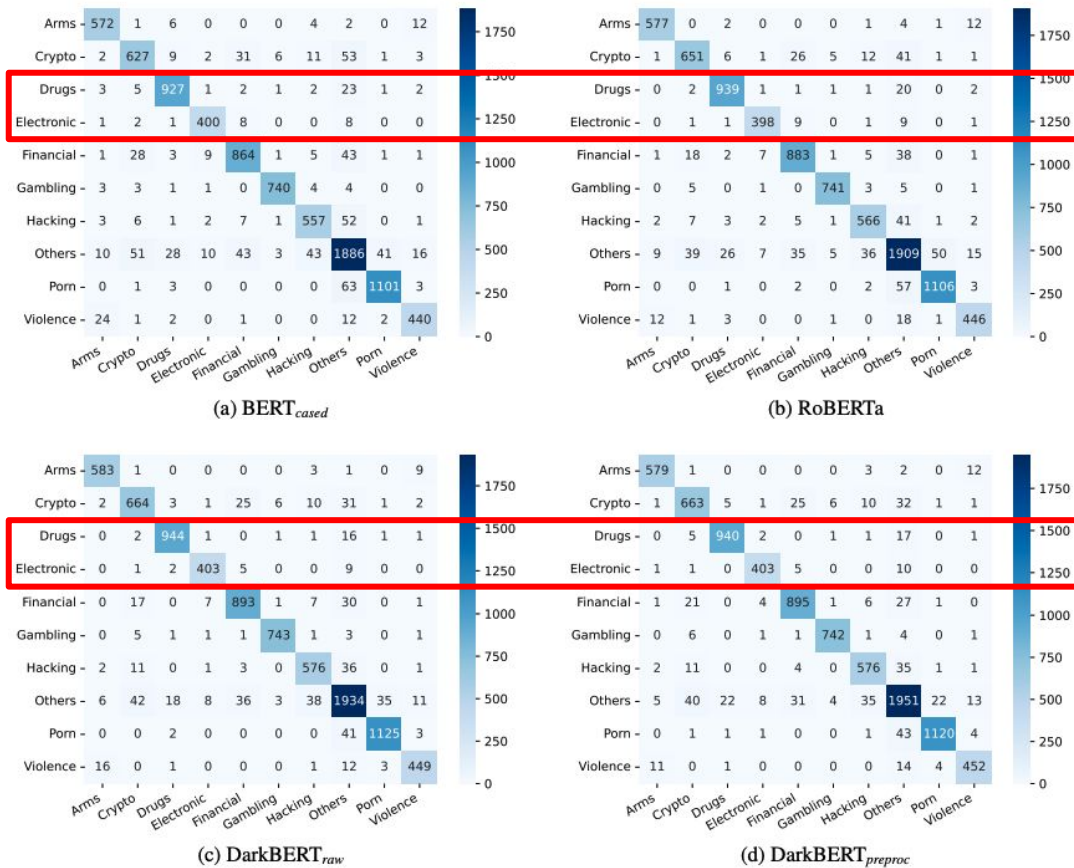


Figure 8: Confusion matrices for selected language models evaluated on the CoDA_{cased} dataset

Use Cases of DarkBERT

1. Ransomware Leak Site Detection

Task : Binary classification of whether a **given dark web site is a ransomware leak site**

Datasets

- Positive data : 150 pages of 54 Ransomware leak site in Dark Web
- Negative data : 679 pages with content similar to that of leak sites (ex, Hacking forum site)
- Train : Validation = 80% : 20%

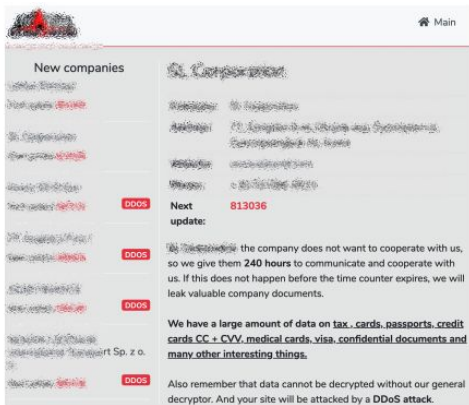


Table 4: Ransomware leak site detection performance. Boldface indicates the best performance.

Input	Model	Precision	Recall	F1 score
Raw	BERT _{cased}	75.83	69.52	71.01
	BERT _{uncased}	77.18	73.90	72.77
	RoBERTa	39.83	36.00	36.27
	DarkBERT_{raw}	78.81	83.62	79.98
Preprocessed	BERT _{cased}	76.81	68.19	70.13
	BERT _{uncased}	71.97	71.62	70.77
	RoBERTa	48.36	45.14	44.31
	DarkBERT_{preproc}	85.16	84.57	84.11

Use Cases of DarkBERT

2. Noteworthy Thread Detection

Task : Binary classification of whether a **given post within a hacking forum is an important post**

Datasets

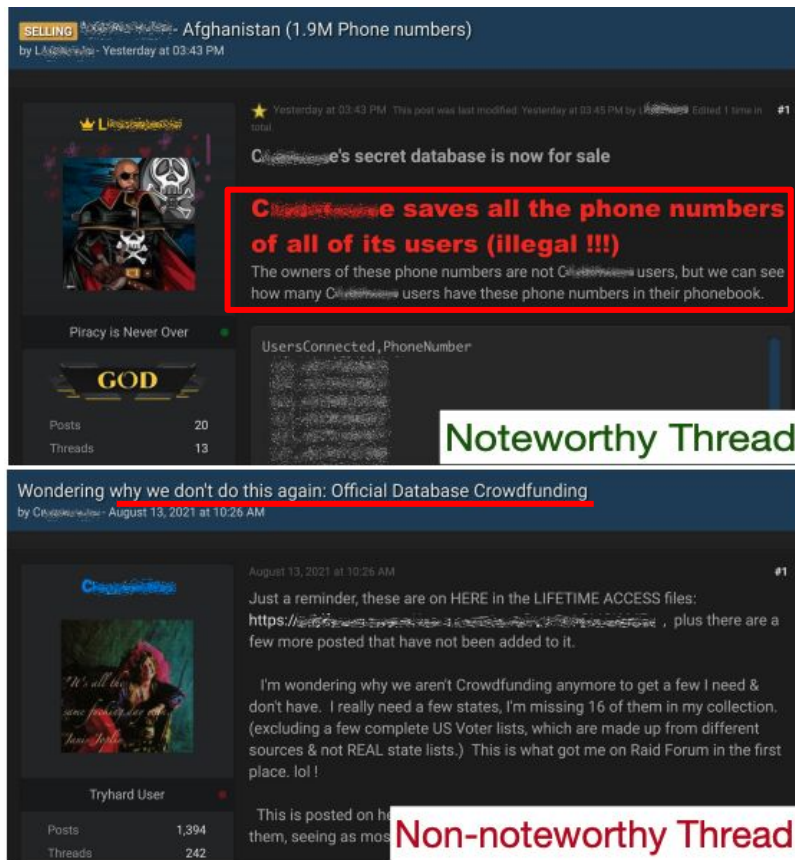
- 1,873 posts of **Raidforums, Breached** Hacking forum sites. (2021.07 ~ 2022.03)
- Criteria of “noteworthy” forums
 1. Sharing of **confidential company assets** ex) admin access, blueprints, source codes
 2. Sharing of **sensitive, private information of individuals** ex) credit info, passports, id, ..
 3. Distribution of **critical malware or vulnerabilities targeting Software or Organizations.**
- 249 positive (important posts) vs 1,624 negative (general posts)
- Train : Validation = 80% : 20%

Use Cases of DarkBERT

2. Noteworthy Thread Detection

Table 5: Noteworthy thread detection performance. Boldface indicates best performance.

Input	Model	Precision	Recall	F1 score
Raw	BERT _{cased}	55.09	19.91	26.90
	BERT _{uncased}	52.34	23.49	28.51
	RoBERTa	28.97	17.89	21.38
	DarkBERT _{raw}	75.93	43.08	52.85
Preprocessed	BERT _{cased}	61.43	20.48	28.81
	BERT _{uncased}	45.46	21.52	26.16
	RoBERTa	29.04	15.27	18.71
	DarkBERT _{preproc}	72.44	45.13	54.17



Use Cases of DarkBERT

3. Threat Keyword Inference

Task

1. Masking drug names from sentences.
2. Verify that the words predicted by the language model are indeed drug-related.

Datasets

- Drug-related text and drug slang datas on Reddit (Zhu et al)

Setting

- Models : DarkBERT_CoDA, BERT_CoDA, BERT_Reddit

fine-tuned on CoDA drugs-classified subset

fine-tuned on Reddit drugs subset



Use Cases of DarkBERT

3. Threat Keyword Inference

- Experiment

CASE EXAMPLE : 25 X XTC 230 MG DUTCH <MASK> PHILIPP PLEIN



Philipp Plein
T-shirt Round Nec...



3axis
Philipp Plein Logo ...



Language Model	Semantically Related Words
DarkBERT	pills , import, md , dot, translation, speed, up, oxy , script, champagne
BERT _{Reddit}	##man, champion, singer, rider, driver, sculptor, producer, manufacturer, ##er, citizen

→ more specific words relevant to drugs

Use Cases of DarkBERT

3. Threat Keyword Inference

- **Results**

DarkBERT shows better performance **when k is small**

- **Discussion**

Despite of DarkBERT's out-performance, results are relatively low due to the limitations of the dataset.

ex) Tesla, Champagne <- Classified as **False Positive**
crystal, ice <- Classified as **True Positive**

Table 7: Quantitative performance metric of threat keyword inference. Precision at k ($P@k$) is measured with varying k in increments of 10.

	Top-10	Top-20	Top-30	Top-40	Top-50
DarkBERT _{CoDA}	0.60	0.60	0.50	0.42	0.42
BERT _{CoDA}	0.40	0.40	0.50	0.50	0.40
BERT _{Reddit}	0.40	0.45	0.60	0.57	0.52

Limitation

1. Limited Usage for Non-English Tasks

- As DarkBERT is pretrained using Dark Web texts in English.
- Challenge : Build a multilingual language model for the Dark Web domain.

2. Dependence on Task-Specific Data

- Shortage of publicly available Dark Web task-specific data.
- Challenge : further manual annotation or handcrafting of necessary data

Conclusion

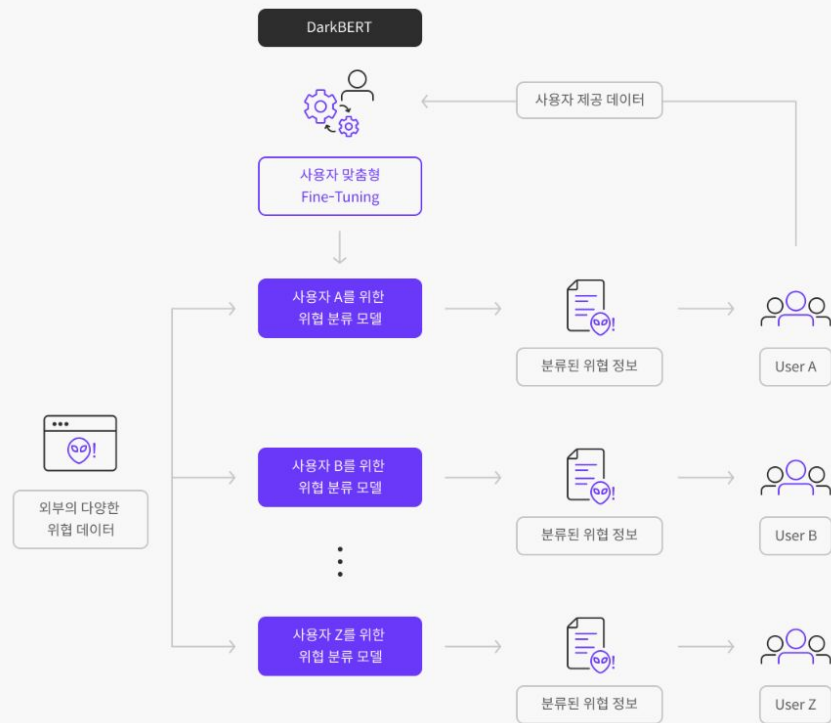
DarkBERT : Dark Web domain-specific language model based by RoBERTa

- Continuous pretraining the model + Preprocessing to adapt well in Dark Web Domain.
- Confident that DarkBERT can contribute a lot to the Cyber Threat industry.

DarkBERT Use Cases in S2W

고객 맞춤형 파인 튜닝과 분류

사용자 맞춤형으로 Fine-Tuning해서 사용할 수 있다.
대량의 다양한 내/외부 비정형 데이터를 처리하여 대량의 데이터로부터
사용자가 원하는 정보만 잘 분류/정제해 줄 수 있다.



DarkBERT Use Cases in S2W

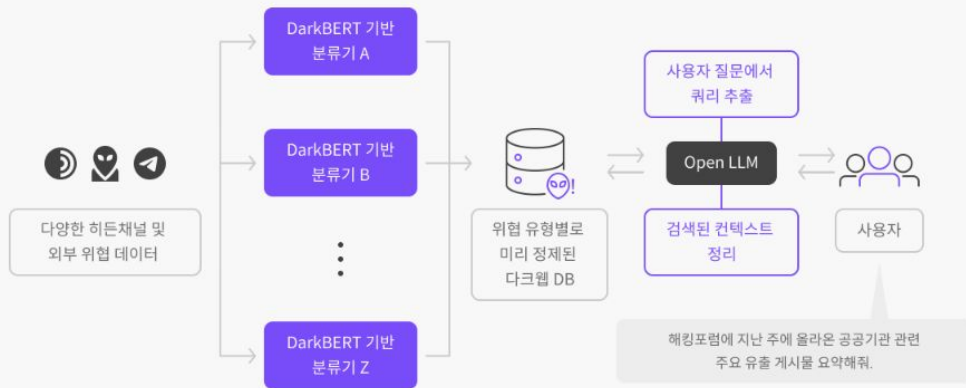
오픈 LLM에 접목

1. 도메인 특화된 데이터 정제/분류

기업의 데이터 특성에 맞게 튜닝된 모델을 이용하여 의사결정에 중요한 데이터를 미리 자동으로 분류하고, 검색 정확도를 높임으로써 LLM 답변 퀄리티를 향상

2. 도메인 특화된 Vectorization for Embedding

RAG의 중요한 요소인 '의미 기반 검색' 수행을 위해선 올바른 의미가 반영된 임베딩이 중요하다. DarkBERT같이 도메인에 특화된 튜닝을 거친 모델은 해당 도메인에서의 의미를 올바르게 반영한 임베딩을 가능케 한다.



DarkBERT Use Cases in S2W

다크웹 특화 생성형 AI : DarkCHAT in XARVIS

demo - 0:27

DarkCHAT

XARVIS라는 다크웹 모니터링 솔루션에 설치된 genAI 모델

실시간으로 수집된 다크웹 데이터에 기반해 현재 다크웹에 어떤 일이 일어나고 있는지 생성한 다크웹 정보를 제공한다.

