Extracting information from darknet market advertisements and forums

- Sven Schlarb  Sven.Schlarb@ait.ac.at, Mina Schütz  Mina.Schuetz@ait.ac.at, Clements Heistracher  clemens.heistracher@ait.ac.at
AIT Austrian Institute of Technology, Competence Unit Data Science & Artificial Intelligence

- Faisal Ghaffar  FAISALGH@ie.ibm.com
IBM Innovation Exchange, Cognitive Computing Group

SECURWARE 2020 November 21, 2020 to November 25, 2020 - Valencia, Spain
Overview

Named entity recognition & Relationship extraction
Manual annotation
Model creation

![Diagram showing Model Creation process]

**Data Set Repository**

- **Manual Annotation**
- **Model Creation**
- **Named Entity Recognition**
- **Relationship Extraction**

**Start Processing**

Processing: grams-entities-model

Selected data sources:
- Data source label: [value]
- Data source process id: [value]
- Representations:
  - [details]

Processing pipeline:
- [named_entity_extraction]

Start processing

[Image of web interface showing processing details]
Working directories

Data set creation
Working directory: 38f02113-9d42-4cf6-be09-517a0942a50f

Data set Repository
Working Directories
Data Set Storage

Annotations & Model Creation
- Manual Annotation
- Model Creation

Information Extraction
- Named Entity Recognition
- Relationship Extraction

Crawler

Repository
- Administration
- Data set creation
- Data set management
- Knowledge Extraction
- Search & Access
- Language
- admin
Named Entity Recognition (CKNER)
Working directories
Process Flow
Process Flow

Harvesting → Scraping → Annotation → Model Creation → Information Extraction

Label suggestions
Annotation interface
(RecogitoJS integrated)
Annotation interface – Loading data

Annotation

dnn-archives-2013-2015-grams-wd

Train Test changed

0 - ALL KINDS OF WEAPONS LONG GUNS TO S
  2. SWISS QUALITY WEED WHITE RUSSIAN Dix
  3. Pistols Gun 6.35 I can put reality photo.
  4. Specifications: Manufacturer: Glock Mo..
  5. This is a custom listing for f***
  6. used cobra .380. Firing conditions, sold.
  7. GR Fusilakamou Genetics: AK-47 selectio..
  8. Dinafem Seeds White Widow is the most po.
  9. NEW Glock 17 4th generation Only 1 avail.
  10. New Glock 37mm pistol never used I hav.
  11. ALL KINDS OF WEAPONS LONG GUNS TO
  12. Activate your Windows 10 with real key f
  13. Variety Ak 47 THC 15 22 indoor Leafly Fo.
  14. Cobra Big Bore Derringer .38 Special M.
  15. New In box Silencer Co SAKER 556 5.56 Sup.
  16. Let us introduce AK47 a staple in canna.
  17. ALL KINDS OF WEAPONS LONG GUNS TO
  18. Taurus Model: 1911 Commander Caliber...
  19. Allright this is the first gun listed.
  20. This is a custom listing for dammo187 e.

Version 1.0 (28.02.2020)
Load grams weapons/drugs dataset
Navigate through annotated postings

Annotation

- You are annotating entities: Weapons/drugs

- AK-47 selection: Smell orange citrus super skunk THC 22.33
  Christmas tree type plant with a characteristic super sticky skunk smell. Tutankhamon has lots of sticky buds with an incredibly high THC level. Starkes NEW super im Geschmack und Konsistens. Es ist auch bestens getrocknet.

Version 1.0 (28.02.2020)
Choose vocabulary for annotation
Create regex pattern (e.g. Regex101.com)
Get named entity suggestions
Regular expression batch labelling
Multi-user
Definition-based batch labelling

- Useful for initial labelling (before user annotation starts)
- Definition file in JSON format required
- Regex terms can be used instead of fixed strings to match tokens
- Filename convention: prelabel_definitions_*\.json

```json
{
    "ACCOUNT": ["account", "address", "bruteforce", "brute force", "username"],
    "BOTNET_DOS": ["botnet", "booter", "ddos", "denial of service"],
    "CREDIT_CARD": ["creditcard", "carding", "ccv", "fullz", "skimmer", ...],
    ...,
    "VULNERABILITY": ["vulnerability", "exploit", "code signing", "windows shell"]
}
```
Definition-based batch labelling

This listing is for 2 operators manuals for two different models of ATMs that are made by the same company. These ATMs are sold all over the world but since I have never left the USA I can not guarantee that they are in your country as I have not been there to see them myself in person. I do know they are in all 56 states even the shitty little town I grew up in. The manuals contain the default master passcodes that if they remain unchanged by the installer must do so for convenience then you can log in as the programmer and change the denomination of the bills that the machine believes it contains. So for instance if you change the denomination of a $20 bill pull out 20 and then you will get 20 $10 bills, because it thinks it is splitting out 4 $20 bills instead of 20 bills and 20 x 1 = 20 to the machine but you really get 200 for 20. I hope fully that makes sense. If you buy this please have a pgp key posted so that I can encrypt the download instructions. If you do not have one posted I will not complete the transaction. This is not a hard thing to do nor do I think it is too much to ask.
Two basic functions for handling train/test data are offered: 1) merge test into train, 2) split train into train and test.

1.) Merge test set into training set
    → Recommended before starting user annotation

2.) Split training set into training and test set
    → Recommended before starting model creation
Merge test set into training set
Split training set into training and test set.
Restore backups
Persist annotations

• Update training/test data files with user annotations (user annotations are removed afterwards)
Technical background:
Named entity recognition
Preparation: Text classification

• Evaluate whether pre-trained embeddings improve the classification performance over simple vector space models, such as TF-IDF, when being transferred to the darknet domain.

• Best performance was achieved with Tensorflow Universal Sentence Encoder and a linear SVM

Heistracher, C. & Schlarb, S.
Machine Learning Techniques for the Classification of Product Descriptions from Darknet Marketplaces
Proceedings of the 11th International Conference on Applied Informatics, 2020
Named entity recognition with SpaCy

• SpaCy is an open-source library for NLP
• SpaCy’s standard models can be used as a basis to build domain-specific models
• For example, SpaCy’s “en_core_web_lg” is a convolutional neural network trained for multiple tasks on the OntoNotes 5 dataset.
• 18 entity types that are available in the pretrained model.
• Words are represented as GloVe vectors, that were trained on the common crawl dataset.
Spacy Named Entity Recognition - Evaluation

- Dataset used for the evaluation was a subset of (Gwern 2015) that contained strings from a list of weapon related terms.

- This subset was manually categorized and the categories were used as labels related to weapons.

- The subset of the data contained 1580 product descriptions including: 1582 Drugs, 769 Weapons, 319 ebooks, 96 3D-printing.

<table>
<thead>
<tr>
<th>Precision</th>
<th>76.97</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>70.88</td>
</tr>
</tbody>
</table>
Technical background: Relation extraction
Relation extraction

- **Focus on Transformer models**
  - Can be pre-trained on smaller datasets
  - Model can be used for NLP downstream tasks

- Several papers already published with BERT, ALBERT, RoBERTA, GPT..

- Mostly trained and evaluated on SemEval 2010 Task 8 & TACRED datasets

- Various approaches
  - Most commonly used: sentence-level
  - Semantic Role Labeling
  - Joint Approaches (NER + RE)
  - Multi-Relation-Extraction
Relation extraction – approach

- Entities are already annotated
- No ground truth for relations for our dataset annotated
  1. Unsupervised relation extraction
  2. Using an already labeled dataset with pre-defined labels such as SemEval 2010 Task 8
- Two models are fitting for our approach:
  - R-BERT
    - BERT, RoBERTa and ALBERT are available
    - Sentence-level
  - OpenNRE
    - BERT
    - Sentence-, bag-, document-level and few-shot
Relation extraction – R-BERT
Relation extraction – OPENnre

NER
Ernest Miller Hemingway was an American journalist

Sentence-level RE
Ernest Hemingway was raised in Oak Park, Illinois

Bag-level RE
In 1921, Ernest Hemingway married Hadley Richardson, the first of his four wives.

Document-level RE
Mark Twain and Olivia Langdon corresponded throughout 1868. She rejected his first marriage proposal, but they were married in Elmira, New York in February 1870. Then, Twain owned a stake in the Buffalo Express newspaper and worked as an editor and writer. While they were living in Buffalo, their son Langdon died of diphtheria at the age of 19 months. They had three daughters: Susy Clemens, Clara Clemens, and Jean Clemens.

Figure 1: The examples of all application scenarios in OpenNRE.
SOURCES


• Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. [https://github.com/monologg/R-BERT](https://github.com/monologg/R-BERT)
Thank you!
IBM’s Moment Recognizer (MoRec)

Faisal Ghaffar (IBM), Brian Maguire (IBM)

Wednesday 19th August 2020
2. Component Overview

- AS an (LEA) end-user, I would like to extract knowledge from DNM forums to
  - understand thread discussions
  - determine topics of discussion
  - determine trends in topics over time
  - identify experts on specific topic(s)
Events timeline

Given a thread, extract who said what over a daily timeline

Hi I don’t have a headphone switch in the volume control, it means that the headphones on my laptop and the speakers work at the same time. I didn’t have this problem with ubuntu 8.10. I really hope there is a solution to the problem because the menu threads I’ve seen on the subject didn’t have a solution, thanks.

Anyone?????? I’ve posted it a while ago but still have not found an answer.

Have you tried adding the head phone volume control? There’s some option to select other volume controls, I’m not on Linux ATM, so try to see if it works.

I don’t have a problem with the headphone volume control, the headphone work fine, the problem is they work together with the speakers when I mute the speakers so does the headphone mute with them. I need the switch that will control which device I am using at the moment.

Seriously? I’m pretty sure my ubuntu has controls separated out. I am not at my ubuntu now, so I can’t tell you the wording, but when I click on the volume applet, it brings up a little slider. On that screen, there’s button that if you click it you can control it individually. Not perfect, but I believe it works, and was there as of this morning! Hope that helps and I’m not just crazy.

It didn’t work, I didn’t even find it, so I created one, rebooted and nothing changed. The volume of the front controls the front and the headphones, and the same with the headphones volume which controls both the headphones and the front speaker.
Complete Process

- Domain Knowledge
  - Topic Detection
  - Word Segmentation
- Domain Dictionary
  - Summarization
  - Post Segmentation
  - Q&A and NE
  - Thread Segmentation
- Preprocessing
- Extracted Text

Events Generation
1. Given a darknet forum, determine topics of discussion from individual post messages.
Text Summarization

Provide ‘extractive’ and ‘abstractive’ summary
Question answering

Given a forum, retrieve answers to questions related to text

**Passage Sentence**

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

**Question**

What causes precipitation to fall?

**Answer Candidate**

gravity

- Between question and answer
- cause---gravity
- precipitation---gravity
- fall---gravity
- what---gravity

Example: question-answer from a context text. Image Source: https://rajpurkar.github.io/reading-comprehension
Model and its Performance

RoBERTa
- Based on BERT with following differences
- Dynamic word-masking instead of static
- Removed NSP training requirement
- Trained on 8k sequences compared to 256 of BERT

Evaluation Metric
• Label Ranking Average Precision
• Final Model : {'LRAP': 0.967, 'eval loss': 0.185}
Component functions

• Multi-Label Categorization based on trained LM
  - Is forum post belong to CaaS and carding, hacking, trafficking, drugs
• Trained Language models (LM) (RoBERTa) on dark net data
• Summarization updated based on LM
• Topic detection based on embeddings
• similarity search function
• Timeline creation from forum threads
• Question-answering (TBD)
• Trade Network extraction and analysis (TBD)
Component Interface (Swagger)

COPKIT MoRec APIs

MoRec component APIs for NLP on DarkNet Forums.

**default**

<table>
<thead>
<tr>
<th>Method</th>
<th>Route</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GET</td>
<td>/ Root</td>
<td></td>
</tr>
<tr>
<td>POST</td>
<td>/api/token_tagger</td>
<td>Token Tagger</td>
</tr>
<tr>
<td>POST</td>
<td>/api/label_posts</td>
<td>Label Post</td>
</tr>
<tr>
<td>POST</td>
<td>/api/tag-all-labels</td>
<td>Token Tagger And Sequence Classifier</td>
</tr>
<tr>
<td>POST</td>
<td>/api/text-summarization</td>
<td>Text Summary</td>
</tr>
<tr>
<td>GET</td>
<td>/api/extract-metadata</td>
<td>Extract Metadata</td>
</tr>
<tr>
<td>POST</td>
<td>/api/plot_events_timeline</td>
<td>Plot Timeline</td>
</tr>
</tbody>
</table>
Thank you!

Time for Questions and Discussion