DreamDrug - A crowdsourced NER dataset for detecting drugs in darknet markets

Johannes Bogensperger
AIT Austrian Institute of Technology
JBogen@gmx.at

Sven Schlarb
AIT Austrian Institute of Technology
Sven.Schlarb@ait.ac.at

Allan Hanbury
TU Wien
allan.hanbury@tuwien.ac.at

Gabor Recski
TU Wien
gabor.recski@tuwien.ac.at

Abstract

We present DreamDrug, a crowdsourced dataset for detecting mentions of drugs in noisy user-generated item listings from darknet markets. Our dataset contains nearly 15,000 manually annotated drug entities in over 3,500 item listings scraped from the darknet market platform "DreamMarket" in 2017. We also train and evaluate baseline models for detecting these entities, using contextual language models fine-tuned in a few-shot setting and on the full dataset, and examine the effect of pretraining on in-domain unannotated corpora.

1 Introduction

This paper describes the construction of DreamDrug, a dataset of drug-related product listings from darknet marketplaces with human-annotated drug entities, suitable for training and evaluating named entity recognition (NER) systems. We provide a detailed description of all steps of annotation and dataset construction, and we also optimize and evaluate standard NER architectures on the new corpus. Our experiments include training models in a few-shot setting and using the full dataset, as well as a study of the effect of domain-adaptive pre-training using unannotated datasets of similar topic and genre. All software used in producing DreamDrug and for reproducing our experiments is published under an MIT license along with sample data files[1], the full dataset is available upon request[2]. Our main contributions are the following:

1. Specifications of an annotation task for marking mentions of drug entities in darknet market listings
2. A detailed description of preprocessing and filtering steps used to create the annotation input
3. The novel NER corpus DreamDrug, annotated by crowdworkers based on these specifications and reviewed manually by the authors
4. Quantitative evaluation of standard supervised learning architectures for NER on the DreamDrug dataset, in a few-shot setting and using the full dataset
5. Three unannotated datasets for domain-adaptive pretraining and experiments demonstrating their impact on the performance of the NER baselines

The paper is structured as follows. In Section 2 we introduce the task and provide an overview of recent related work on information extraction from noisy user-generated text. Section 3 describes the corpus construction process. Section 4 and Section 5 present our baseline experiments and results, respectively, and Section 6 provides a brief conclusion.

2 Background

Extracting information from text concerning illicit drug consumption or trade is a non-trivial task, since such information is rarely available in a structured form. Patients who misuse their medication are more likely to discuss this on social media than with medical professionals (Bigeard et al., 2018), while drug dealers intentionally misspell names of drugs on social media to prevent automatic detection (Li et al., 2019). Drug-specific named entity recognition (NER) models are most often developed for the medical domain and cannot handle such misspellings, the usage of slang, and other characteristics of noisy user-generated texts such as lack of punctuation and non-standard grammar (Rezaei et al., 2020). Many of these issues are

[1] https://github.com/jbogensperger/DRUG_CROSSNER

[2] DreamDrugDataset@gmail.com

Proceedings of the 2021 EMNLP Workshop W-NUT: The Seventh Workshop on Noisy User-generated Text, pages 137–157
November 11, 2021. ©2021 Association for Computational Linguistics

peruanse 92 cocain very good snife and for people tghe want cooking out very good smills good frech from the block orginail good good good AAA.

Figure 1: Sample listing from the *DreamMarket* dataset

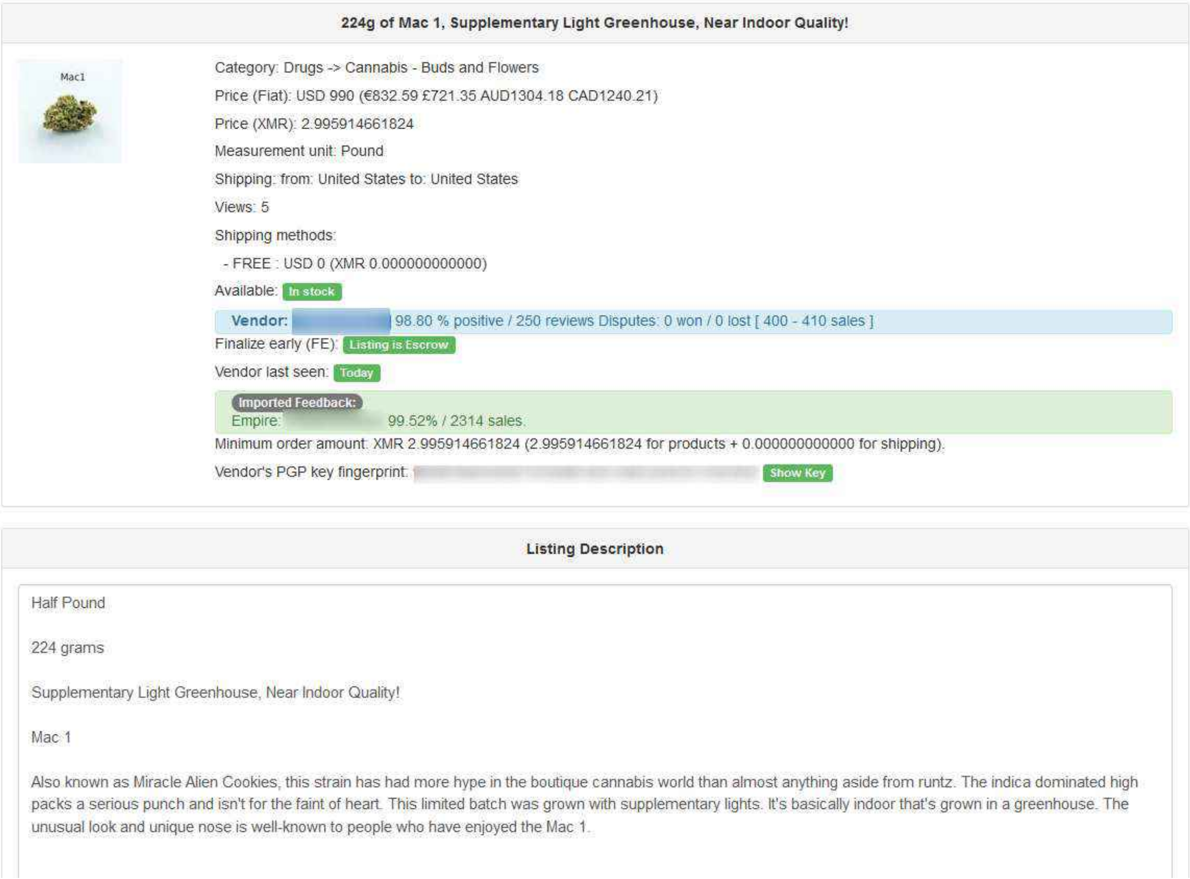
illustrated by the example listing from a darknet marketplace in Figure 1.

Still, many organizations are dependent on extracting information about drugs or similiar goods from noisy online sources. Public health organizations study drug use patterns in society such as the opioid epidemic in the United States (Ostling et al., 2018), pharmaceutical companies can collect evidence of adverse drug reactions from social media data (Chen et al., 2017), and law enforcement agencies can track illicit trading activities in darknet markets. Each of these use-cases can be supported by the automatic detection of drug entities in noisy user-generated text. To enable the development of such systems we create a dataset for training and evaluation by annotating drug-related item listings of darknet marketplaces. An example of such a listing is shown in Figure 2.

2.1 NER in User-Generated Texts

Much of recent research on named entity recognition in user-generated texts was facilitated by the 2017 W-NUT Shared Task on Novel and Emerging Entity Recognition (Derczynski et al., 2017). The dataset provided by the organizers of the competition was compiled from a variety of social media sources (Twitter, Reddit, Youtube, StackExchange), annotated for 6 categories (person, location, corporation, product, creative work, group) by multiple crowdworkers, and corrected by the authors. The difficulty of the task of detecting novel entities in noisy domains is indicated by the relatively low performance of all competing systems. The top-performing system of Aguilar et al. (2017) achieved an average F-score of 41.86 on the evaluation dataset. Their system combines character- and word-level features with gazetteers for each entity type, and uses them as input to a Bidirectional Long Short-term Memory (BiLSTM) network with a Conditional Random Fields (CRF) classifier. Other top-performing approaches included Jansson and Liu (2017), who use LDA topic modeling for creating word-level features for an LSTM classifier with a CRF output layer, and Lin et al. (2017), who incorporated part-of-speech tags and dependency relations in their word representations and also used a BiLSTM with CRF output layer. A significant improvement on this benchmark was achieved by Akbik et al. (2019b) using pooled contextualized embeddings, a method to aggregate character-level representations from contextual language models. This method is implemented in the FLAIR framework (Akbik et al., 2019a) and achieves an F-score of 49.59 on the W-NUT 2017 dataset.

We are aware of two annotated NER corpora from the darknet domain. Durrett et al. (2017) created a dataset for identifying products in cybercrime marketplace postings, annotating 1,938 posts across 4 forums. When training NER models on data from the same forum, they achieve F-score performances between 83 and 90, in a cross-forum setting their performance drops by 7-11 points. A more recent darknet dataset is NUToT (Nabki et al., 2020), containing 851 sentences from darknet sources annotated manually with the categories of the W-NUT challenge. Their experiments involve compensating for the absence of domain-specific gazetteers with Local Distance Neighbor (LDN) features, a method for enriching word representations with nearest neighbors in a word embedding space. Performance of top-performing systems on their dataset is in the range of 40 to 60 across categories (F-scores of entities). The corpus we introduce in this paper is annotated for a single entity type (drug) that is more specific than the categories of W-NUT and NUToT and allows our simple baselines to achieve F-scores above 70 even in a few-shot setting and above 80 when using the full dataset (see Section 5). In addition to enabling higher accuracy, we believe that NER datasets that correspond directly to a narrow information extraction task are better models of real-world NER use-cases than the commonly used datasets labeled for broad categories such as *person*, *location*, *organization*, or even *product*. While the fine-grained task definiton necessary for consistent annotation of DreamDrug (see Section 3.2) requires considerations specific to a single task, it also provides an example of the narrow and domain-specific entity categories that are typical of real-world NER applications. Besides providing a novel dataset and baseline models, our work also demonstrates that entity recognition tasks for narrow domains and specialized text genres can benefit from even small amounts of annotated data.



224g of Mac 1, Supplementary Light Greenhouse, Near Indoor Quality!

Category: Drugs -> Cannabis - Buds and Flowers

Price (Fiat): USD 990 (€832.59 £721.35 AUD1304.18 CAD1240.21)

Price (XMR): 2.995914661824

Measurement unit: Pound

Shipping: from: United States to: United States

Views: 5

Shipping methods:

- FREE : USD 0 (XMR 0.000000000000)

Available: In stock

Vendor: [redacted] 98.80 % positive / 250 reviews Disputes: 0 won / 0 lost [400 - 410 sales]

Finalize early (FE): Listing is Escrow

Vendor last seen: Today

Imported Feedback:
Empire: [redacted] 99.52% / 2314 sales.

Minimum order amount: XMR 2.995914661824 (2.995914661824 for products + 0.000000000000 for shipping).

Vendor's PGP key fingerprint: [redacted] Show Key

Listing Description

Half Pound

224 grams

Supplementary Light Greenhouse, Near Indoor Quality!

Mac 1

Also known as Miracle Alien Cookies, this strain has had more hype in the boutique cannabis world than almost anything aside from runtz. The indica dominated high packs a serious punch and isn't for the faint of heart. This limited batch was grown with supplementary lights. It's basically indoor that's grown in a greenhouse. The unusual look and unique nose is well-known to people who have enjoyed the Mac 1.

Figure 2: Sample drug item listing on the darknet platform *White House Market*

2.2 Crowdsourcing for NER

Crowdsourcing, the practice of soliciting non-expert annotators for assistance in creating datasets on a scale, usually with the help of online platforms, has become a standard practice of researchers aiming to build annotated corpora with limited resources. The most common scheme, where crowd workers are rewarded for each annotation task they complete, is sometimes referred to as *mechanized labor* (Sabou et al., 2014). In Section 3.3 we shall give a detailed description of the entire process of creating the DreamDrug dataset. For defining the Named Entity Recognition task we consult several recent sets of guidelines for NER tasks (Finin et al., 2010; Nédellec et al., 2006; Benikova et al., 2014). When designing the annotation process we rely on discussions of annotation methodologies in (Sabou et al., 2014) and (Fort et al., 2009). Approaches to measuring inter-annotator agreement are studied in detail by Asheghi et al. (2014), crowdsourcing techniques are discussed by Feyisetan et al. (2018), and we also rely on the work of Gomes et al. (2020) for initial effort estimation, which we see as essential in ensuring the ethical treatment of annotators.

3 Corpus construction

3.1 Data preparation

The starting point for our corpus construction process is a segment of the AZSecure dataset (Du et al., 2018) containing 91,463 darknet market listings scraped from *Dream Market*, collected between 2013 and 2017. Listing documents typically contain a product name and description along with metadata such as price, quantity, contact information of the vendor, etc. We extracted 45,446 items belonging to categories associated with drugs and performed a series of data cleaning steps. First, deduplication removed all listings containing exactly the same description text as another item listing. The English only step removed all listings with a description text categorized as other than English by the language-detection library (Shuyo, 2010). Next, outlier lengths removed listings with a description text longer than 3000 characters or shorter than 30 characters. While this step also removed some relevant content, it efficiently filtered product listings with long boilerplate sections (e.g. about delivery details), as well as items that contained no relevant text (e.g. a product description might only state *Free shipping*). Finally, `soft deduplication` removed listings with a description text whose first 100 characters contained the same non-numeric characters as another listing's description, characteristic of listings of the same product with different quantities. The resulting dataset contains 11,674 items, Table 1 shows the number and percentage of listings removed by each filtering step.

filtering step	# removed	% removed	# kept
deduplication	25,012	55.0%	20,434
English only	2,102	10.3%	18,332
outlier lengths	1,488	8.1%	16,844
soft deduplication	5,170	30.7%	11,674

Table 1: Number and percentage of item listings filtered in each step (see text for description of each step)

For the construction of the `DreamDrug` dataset we used the product name and product description fields of the 11,674 remaining listings, the final dataset contains 3,507 items randomly selected from this set. Common unique identifiers of vendors in the product description texts were replaced with random values. Telephone numbers were detected using the python library `phonenumber`[3], email addresses and URLs were replaced using regular expressions (see Appendix B) and the `faker`[4] package, while vendor names were available as separate fields in the original database. Finally, product description texts were tokenized using Stanza[5] (Qi et al., 2020), version 1.2.0.

3.2 Task definition

Our definition of which sequences of words constitute drug entities was developed in an initial set of experiments that involved manual annotation of 500 item listings by the authors. Our starting point was the definition used by the Food and Drug Administration (FDA) (U.S. Food and Drug Administration, 2021):

1. *"articles intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease"*

2. *"articles (other than food) intended to affect the structure or any function of the body of man or other animals."*

The guidelines we developed for consistent annotation clarify that for an entity to be labeled it must uniquely identify a specific article according to the FDA definition. Besides direct mentions of substances (*cocaine, THC*, etc.), slang terms that clearly identify a drug, such as *speed* and *pep* for amphetamine or *Mary Jane* for cocaine, are also considered drug entities. Phrases that do not uniquely identify a substance, such as *Charley's mellow Sleeper Bars* or *Green Dom Perignon*, are not to be annotated. Following the FDA definition we do not annotate substances occurring naturally in the human body such as *testosterone* or *dopamine*. Phrases referring to the physical form of a drug only are also not annotated, even if the form is unique to a particular drug, as e.g. *blotter* is to LSD. The common term *mushroom* constitutes a corner case, as one may argue that it refers to the form of a drug, but since in this dataset it clearly identifies a particular chemical (psilocybin, the active agent of all types of *magic mushrooms*), we decided to annotate it as a drug entity. Some phrases describing drugs contain multiple elements that could each be annotated as a separate drug entity based on our definition. Examples include phrases such as *Viagra Sildenfanil* and *Haze weed* — we annotate these as two subsequent drug spans. The annotation guidelines provided to crowd workers contain a simplified description of our definition with examples and is presented in Appendix C. All documents annotated by crowd workers were subsequently reviewed and corrected manually by the authors to ensure high quality and consistent treatment of corner cases — details of the annotation and review process are presented in Section 3.3.

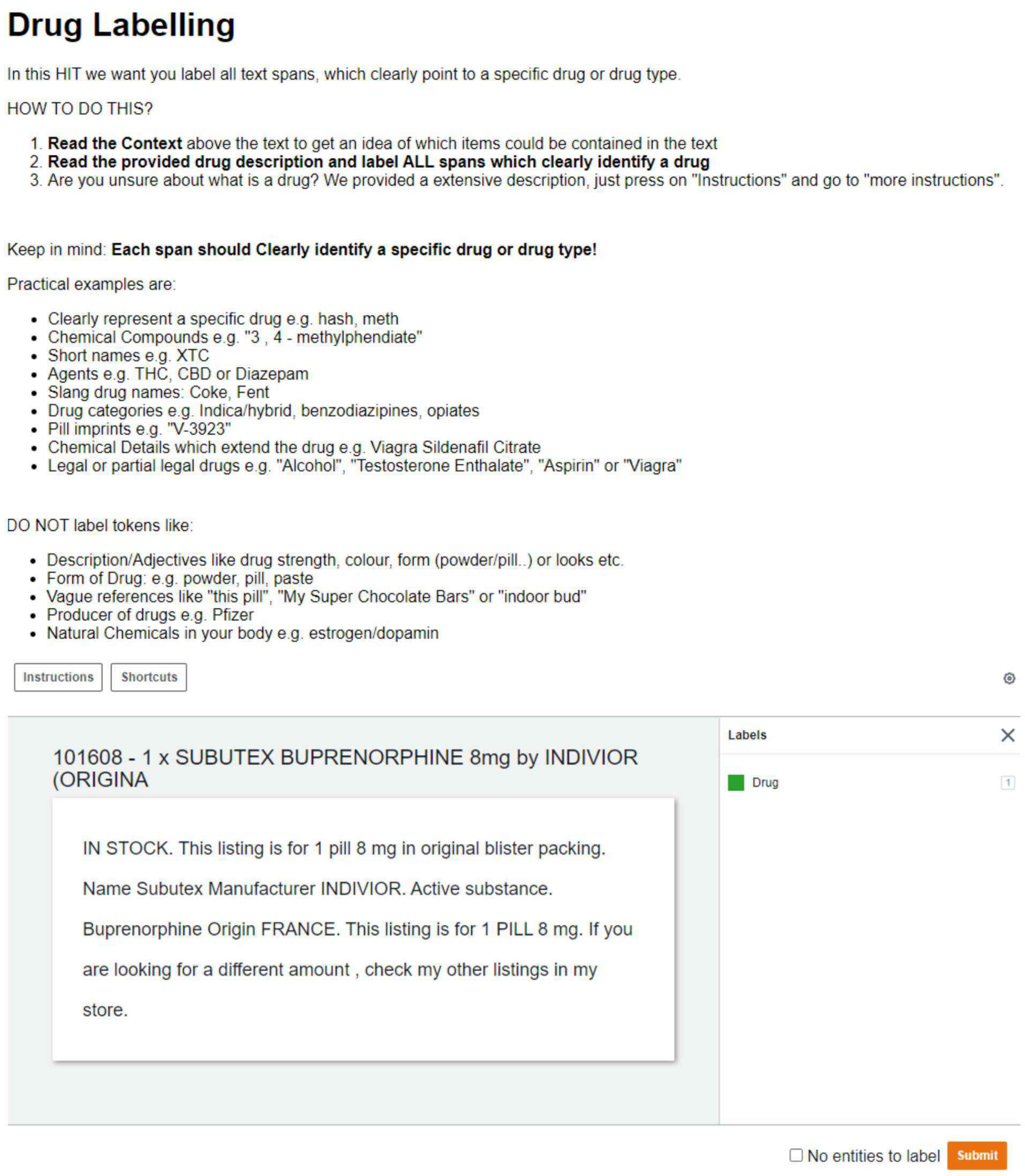
3.3 Data annotation

Crowdsourced annotation was performed using the definition of drug entities presented in Section 3.2. The guidelines presented to annotators were revised several times during pilot experiments that involved the implementation of two annotation interface prototypes using the platforms Appen and Amazon Mechanical Turk (AMT), both of which are regularly used for annotating NER datasets (Jalal et al., 2020; Feyisetan et al., 2015; Bontcheva et al., 2017; Feyisetan et al., 2018). Appen was then chosen, primarily because of its built-in capabilities for evaluating annotator performance and for screening annotators using test questions. Due to an unexpected

[3] `https://pypi.org/project/phonenumbers/`

[4] `https://github.com/joke2k/faker`

[5] `https://stanfordnlp.github.io/stanza/`



Drug Labelling

In this HIT we want you label all text spans, which clearly point to a specific drug or drug type.

HOW TO DO THIS?

1. **Read the Context** above the text to get an idea of which items could be contained in the text
2. **Read the provided drug description and label ALL spans which clearly identify a drug**
3. Are you unsure about what is a drug? We provided a extensive description, just press on "Instructions" and go to "more instructions".

Keep in mind: **Each span should Clearly identify a specific drug or drug type!**

Practical examples are:

- Clearly represent a specific drug e.g. hash, meth
- Chemical Compounds e.g. "3 , 4 - methylphendiate"
- Short names e.g. XTC
- Agents e.g. THC, CBD or Diazepam
- Slang drug names: Coke, Fent
- Drug categories e.g. Indica/hybrid, benzodiazipines, opiates
- Pill imprints e.g. "V-3923"
- Chemical Details which extend the drug e.g. Viagra Sildenafil Citrate
- Legal or partial legal drugs e.g. "Alcohol", "Testosterone Enthalate", "Aspirin" or "Viagra"

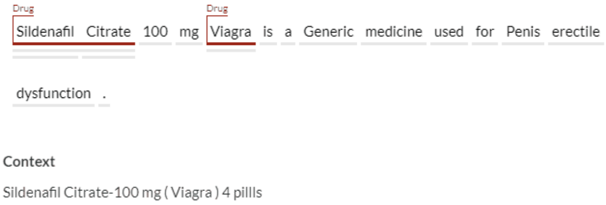
DO NOT label tokens like:

- Description/Adjectives like drug strength, colour, form (powder/pill..) or looks etc.
- Form of Drug: e.g. powder, pill, paste
- Vague references like "this pill", "My Super Chocolate Bars" or "indoor bud"
- Producer of drugs e.g. Pfizer
- Natural Chemicals in your body e.g. estrogen/dopamin

101608 - 1 x SUBUTEX BUPRENORPHINE 8mg by INDIVIOR (ORIGINA

IN STOCK. This listing is for 1 pill 8 mg in original blister packing. Name Subutex Manufacturer INDIVIOR. Active substance. Buprenorphine Origin FRANCE. This listing is for 1 PILL 8 mg. If you are looking for a different amount , check my other listings in my store.

Figure 3: Screenshot of the annotation interface on Amazon Mechanical Turk. The long version of the annotation guidelines (see Appendix C) was accessible via the *Instructions* button.



Sildenafil Citrate 100 mg Viagra is a Generic medicine used for Penis erectile dysfunction .

Context

Sildenafil Citrate-100 mg (Viagra) 4 pillls

Figure 4: Screenshot of the annotation interface on Appen.

IAA	AMT	Appen
Cohen's Kappa	0.76	0.43
Micro F1	0.79	0.55
Macro F1	0.78	0.60

Table 2: Inter-Annotator Agreement measures for the overall annotations by crowd workers of Amazon MTurk and Appen. Macro F1 is the average of F1 scores across pairs of annotators.

	MTurk	Appen
% of characters added	8.56	25.81
% of characters deleted	3.51	0.89
% of spans added	13.19	27.49
% of spans deleted/altered	8.65	10.45

Table 3: Average effort of the review process.

licensing issue we were forced to switch platforms during the annotation process, and the final dataset was constructed using annotations from both Appen and AMT, agreement figures and dataset statistics in this section are provided for both platforms. Crowdworkers were presented with the text of one listing at a time and asked to mark the beginning and end of sequences of words that identify a drug. The title of the listing was provided as context. Screenshots of the annotation interfaces on Amazon Mechanical Turk and Appen are shown in Figures 3 and 4, respectively. On Appen, instructions were provided in a separate window and contained the full annotation guidelines shown in Appendix C. These longer guidelines were available to AMT workers when they clicked the *Instructions* button.

Annotation quality was monitored by measuring Inter-Annotator Agreement (IAA) and by conducting manual reviews of annotation samples from each worker. Parameters of the annotation process such as reward per task, maximum time per assignment, and number of annotations solicited for each listing, were updated during the annotation process based on our experience — a detailed description of this process is available in the Appendix of (Bogensperger, 2021). We obtained over 4,000 valid annotations from Appen and over 7,400 from AMT. Appen annotators received a fixed payment of either $0.06 or $0.08 per annotated document (the exact rate changing across batches), on AMT rates also varied across annotators and payments per row ranged between $0.09 and $0.30, while hourly rates varied between $12 and $18. Our total incurred costs were $548 on Appen and $1505 on AMT, in the former case this included payment for 995 additional annotations that were invalidated based on annotators' performance on the test questions. Table 2 shows inter-annotator agreement (IAA) values for annotations obtained via each of the two platforms. Following Viera and Garrett (2005) and Hripcsak and Rothschild (2005) we used F1-Agreement as the primary IAA measure, Cohen's Kappa is also calculated for comparison. Spans annotated by at least two annotators were reviewed and corrected manually by the authors, then all annotated documents were reviewed once again by the authors to find missing annotations. The review process followed the principles described in Section 3.2. Table 3 shows the amount of editing performed for each portion of the annotations. Both of these figures and the agreement scores show a considerable difference between the quality of annotations from the two platforms, and without further study we can only speculate that this may be attributed to the continuous improvement of the annotator instructions based on our experiences, the different presentations of these instructions (MTurk annotators had access to a short and long version), or to the different incentive systems of the two platforms (AMT workers were required to hold a "master" qualification on the platform, which they could lose when providing low-quality annotations, while Appen workers' performance was only controlled on the test questions).

The final dataset contains 3,507 item listings with 364,003 tokens, 14,934 annotated spans and 3,048 unique drug entities. The dataset was split into training, validation, and test portions of 2244, 561, and 702 listings, respectively. Table 4 provides basic descriptive statistics about the `DreamDrug` dataset, further statistics as well as a list of the most frequent spans are presented in Appendix A. Manual evaluation of a sample of rare spans reveals that while some of them refer to

	Min	Max	Mean	Median
words / doc	3	534	103.8	61
chars / doc	25	2,930	572.6	334
words / entity	1	20	1.9	2
chars / entity	1	80	12.1	10
entity / doc	0	101	4.26	2

Table 4: Descriptive Statistics of dataset

rare entities, such as names of cannabis strains (e.g. *alien technology*) or specific chemical descriptors (e.g. *n-acetyl-p-aminophenol*), most of them are spelling variations of common drugs (e.g. "Oxycodone hydrochloride" instead of ... "Oxycodone HCL") and misspellings ("Oxymorophone" instead of "Oxymorphone").

4 Experiments

We train a set of baseline transformer-based models, examine the effects of domain-adaptive pretraining using multiple datasets, and also evaluate top-performing models in a few-shot training scenario.

4.1 Baseline models

We train the standard BERT-based architecture for Named Entity Recognition (Devlin et al., 2018), which consists of a transformer layer and a linear layer with dropout, using the PyTorch-based implementations from Huggingface (Wolf et al., 2020), similar to the system in (Liu et al., 2021). We train systems based on the `bert-base-cased` and `roberta-base` models with default parameters, using the Adam optimizer, a learning rate of $5 \cdot 10^{-5}$, and a batch size of 4. For the few-shot learning setup we used only the first 100 documents from the training set. In both setups we trained our models for 10 epochs. We also train the FLAIR[6] system (Akbik et al., 2019a) using default hyperparameters and training for 50 epochs. We use BERT (`bert-base-cased`) as the input representation of the FLAIR system because it yields better results than FLAIR's own contextualized embedding (Akbik et al., 2019b). For evaluation we use standard NER metrics as defined by the CoNLL 2003 shared task (Tjong Kim Sang and De Meulder, 2003), computing precision, recall, and F-score, where a predicted named entity is considered correct only if it is an exact match of a gold standard entity. While Named Entity Recognition is generally implemented on the sentence level, the relatively short average length of listings (see Table 4) allows us to train the baselines on full documents, limited only by the maximum input length of BERT, 512 (subword) tokens. While the average number of words per document in our dataset is 103.79, the average number of subword tokens output by BERT's tokenizer is 152.18, and 4.1% of documents are longer than the maximum input length of 512, these were clipped at 512 tokens for training and evaluation of each of our baselines. A final preprocessing step involved removing all punctuation except commas from item descriptions to reduce noise, the regular expressions we used are presented in Appendix B.

4.2 Domain adaptation

We improve the performance of our baselines by pretraining on unannotated text representing the domain and/or genre of the `DreamDrug` dataset. We compile three datasets that can be used separately or in combination. `Dreammarket` contains Dreammarket product listings of drugs that were not used in creating `DreamDrug`. The size of this dataset is slightly above 800 000 words. The `Grams` dataset is constructed from the Darknet Market Archives (Branwen et al., 2015), using the subsets *Abraxas, Agora, Alpha, ME*, and *Oxygen*, and contains listings from a variety of domains. This dataset contains nearly 1.7 million words. The third dataset used for domain adaptation, `Wikipedia`, was constructed using 3,210 drug-related articles downloaded from Wikipedia[7] using its Python API[8]. Sections of Wikipedia articles that do not typically contain running text (*References*, *See also*, and *External Links*) were removed, and the remaining text was searched for possible mentions of drug entities by annotating each sentence for mentions of entities in the knowledge graph DBPedia (Auer et al., 2007) using the `annotate` function of the Spotlight API[9], which performs fuzzy matching. Drug entities were detected based on the presence of the DBPedia attributes `dbp:legalUs` or `dbp:legalUn`, sentences containing entities with either of these attributes were added to the dataset used for domain adaptation. The size of the resulting dataset is

[6] `https://github.com/flairNLP/flair`

[7] `https://en.wikipedia.org/`

[8] `https://pypi.org/project/wikipedia/`

[9] `https://pypi.org/project/spacy-dbpedia-spotlight/`

Base	DAPT	D	F1	Prec.	Rec.
FLAIR	None	NA	59.72	60.29	59.16
BERT	None	0	63.80	67.84	60.21
BERT	All	0	69.86	72.91	**67.06**
BERT	None	0.5	62.93	70.59	56.77
BERT	DM	0.5	69.43	74.20	65.23
BERT	Grams	0.5	68.49	71.87	65.41
BERT	Wiki	0.5	64.32	68.76	60.42
BERT	All	0.5	**70.98**	**76.73**	66.04
RoBERTa	All	0.35	69.11	73.54	65.18

Table 5: Performance of top baseline configurations in the few-shot scenario. The column DAPT indicates the dataset(s) used for domain-adaptive pretraining, D indicates dropout rate.

more than 880 000 words, the three datasets for domain adaptation together contain nearly 3.4 million words.

5 Results

Tables 5 and 6 present the perfomance of top baseline configurations in the few-shot setting and when using the full training set. The BERT-based systems achieved higher scores in the few-shot setup, while RoBERTa-based setups performed better when trained on the full dataset. For both setups we include the performance of the top system with and without dropout as well as with domain-adaptive pretraining (DAPT) on each of the additional datasets. For both the few-shot setup and the full training set we also include the best-performing configuration of the other language model (RoBERTa for the few-shot setting and BERT for the full dataset) and the performance of the best FLAIR model. The effect of domain adaptation using in-domain unannotated text is especially pronounced in the few-shot setting, where it achieves an 8 point increase in F-score, 6.5 of which can be achieved using only the DM dataset, which is the one most similar to the listings in DreamDrug. We believe this shows the strong potential of pretraining on unannotated text that is characteristic of both the domain and the genre of a novel task, especially in settings where the availability of human-annotated data is limited. Finally, Table 7 contains the results of our top models on the test set.

6 Conclusion

We have constructed the novel dataset DreamDrug for the task of recognizing drug entities in darknet market listings, used it to train and evaluate baseline NER models, and demonstrated the effectiveness of domain-adaptive pretraining on large unannotated datasets. As one of the first entity-annotated corpora of darknet text, and the first task-specific one, DreamDrug will facilitate the study of named entity recognition and related information extraction tasks in noisy user-generated texts. Additionally, it can also be extended with several types of annotation, including entities from multiple classes, relations such as "drug-quantity" or "drug-strength", or links from entity mentions to a knowledge base.

Base	DAPT	D	F1	Prec.	Rec.
FLAIR	None	N/A	81.42	80.83	**82.01**
RoBERTa	None	0	80.84	81.13	80.56
RoBERTa	Wiki	0	80.83	83.65	78.20
RoBERTa	All	0	81.21	**86.63**	76.43
RoBERTa	None	0.1	80.55	81.43	79.69
RoBERTa	DM	0.1	80.41	83.19	77.80
RoBERTa	Grams	0.1	80.52	84.99	76.49
RoBERTa	Wiki	0.1	**82.16**	84.86	79.62
RoBERTa	All	0.1	81.51	85.37	77.98
BERT	All	0.45	81.64	85.71	77.94

Table 6: Performance of top baseline configurations using the full training set. The column DAPT indicates the dataset(s) used for domain-adaptive pretraining, D indicates dropout rate.

Model	F1	Prec.	Rec.
few-shot training			
FLAIR	62.13	65.18	59.35
Our best model	73.55	78.38	69.28
full training set			
FLAIR	81.82	82.75	80.92
Our best model	83.86	84.83	82.92

Table 7: Performance of top baseline configurations on the test set

Acknowledgements

We thank the two anonymous reviewers for their insightful comments and for additional references. The research described in this paper was carried out as part of the COPKIT project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 786687.

Ethical considerations

An analysis of potential project internal and external impacts of the research activity was conducted as part of the COPKIT project. The relevance of data protection principles, such as lawfulness, fairness, and transparency, purpose limitations, data minimization, accuracy, storage limitation, integrity and confidentiality, and accountability has been analyzed for this dataset publication. The activity was led by the project internal ethical, legal, and privacy team. Regarding concrete measures, the dataset was pseudonymized to remove any personal data, and the base dataset for ground-truth creation was approved for use by the ethical, legal, and privacy team. It was decided to provide access to the data only on request to prevent misuse of the dataset for purposes other than research.

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153, Copenhagen, Denmark. Association for Computational Linguistics.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019a. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019b. Pooled contextualized embeddings for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 724–728, Minneapolis, Minnesota. Association for Computational Linguistics.

Noushin Rezapour Asheghi, Serge Sharoff, and Katja Markert. 2014. Designing and evaluating a reliable corpus of web genres via crowd-sourcing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1339–1346, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, page 722–735, Berlin, Heidelberg. Springer-Verlag.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D named entity annotation for German: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).

É. Bigeard, N. Grabar, and F. Thiessard. 2018. Detection and analysis of drug misuses. A study based on social media messages. *Frontiers in Pharmacology*, 9.

Johannes Bogensperger. 2021. Exploring transfer learning techniques for named entity recognition in noisy user-generated text. Master's thesis, TU Wien.

Kalina Bontcheva, Leon Derczynski, and Ian Roberts. 2017. Crowdsourcing named entity recognition and entity linking corpora. In *Handbook of Linguistic Annotation*, pages 875–892. Springer.

Gwern Branwen, Nicolas Christin, David Décary-Hétu, Rasmus Munksgaard Andersen, StExo, El Presidente, Anonymous, Daryl Lau, Delyan Kratunov Sohhlz, Vince Cakic, Van Buskirk, Whom, Michael McKenna, and Sigi Goode. 2015. Dark net market archives, 2011-2015. `https://www.gwern.net/DNM-archives`. Accessed: 25.01.2021.

X. Chen, Myrtille Deldossi, Rim Aboukhamis, C. Faviez, B. Dahamna, P. Karapetiantz, Armelle Guenegou-Arnoux, Y. Girardeau, Sylvie Guillemin-Lanne, A. L. Louët, N. Texier, Anita Burgun-Parenthoine, and S. Katsahian. 2017. Mining adverse drug reactions in social media with named entity recognition and semantic methods. *Studies in health technology and informatics*, 245:322–326.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Po-Yi Du, Ning Zhang, Mohammedreza Ebrahimi, Sagar Samtani, Ben Lazarine, Nolan Arnold, Rachael Dunn, Sandeep Suntwal, Guadalupe Angeles, Robert Schweitzer, and Hsinchun Chen. 2018. Identifying, collecting, and presenting hacker community data: Forums, irc, carding shops, and dnms. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 70–75.

Greg Durrett, Jonathan K. Kummerfeld, Taylor Berg-Kirkpatrick, Rebecca Portnoff, Sadia Afroz, Damon McCoy, Kirill Levchenko, and Vern Paxson. 2017. Identifying products in online cybercrime marketplaces: A dataset for fine-grained domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2598–2607, Copenhagen, Denmark. Association for Computational Linguistics.

O. Feyisetan, Markus Luczak-Rösch, E. Simperl, Ramine Tinati, and N. Shadbolt. 2015. Towards hybrid NER: A study of content and crowdsourcing-related performance factors. In *ESWC*.

O. Feyisetan, E. Simperl, Markus Luczak-Rösch, Ramine Tinati, and N. Shadbolt. 2018. An extended study of content and crowdsourcing-related performance factors in named entity annotation. *Semantic Web*, 9:355–379.

Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88, Los Angeles. Association for Computational Linguistics.

Karën Fort, Maud Ehrmann, and Adeline Nazarenko. 2009. Towards a methodology for named entities annotation. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 142–145, Suntec, Singapore. Association for Computational Linguistics.

Inês Gomes, Rui Correia, Jorge Ribeiro, and João Freitas. 2020. Effort estimation in named entity tagging tasks. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 298–306, Marseille, France. European Language Resources Association.

G. Hripcsak and A. Rothschild. 2005. Technical brief: Agreement, the F-Measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 12 3:296–8.

Mona Jalal, Kate K. Mays, L. Guo, and Margrit Betke. 2020. Performance comparison of crowdworkers and NLP tools on named-entity recognition and sentiment analysis of political tweets. *ArXiv*, abs/2002.04181.

Patrick Jansson and Shuhua Liu. 2017. Distributed representation, LDA topic modelling and deep learning for emerging named entity recognition from social media. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 154–159, Copenhagen, Denmark. Association for Computational Linguistics.

Jiawei Li, Q. Xu, Neal Shah, and T. Mackey. 2019. A machine learning approach for the detection and characterization of illicit drug dealers on Instagram: Model evaluation study. *Journal of Medical Internet Research*, 21.

Bill Y. Lin, Frank Xu, Zhiyi Luo, and Kenny Zhu. 2017. Multi-channel BiLSTM-CRF model for emerging named entity recognition in social media. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 160–165, Copenhagen, Denmark. Association for Computational Linguistics.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. CrossNER: Evaluating cross-domain named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13452–13460.

Mhd Wesam Al Nabki, E. Fidalgo, E. Alegre, and Laura Fernández-Robles. 2020. Improving named entity recognition in noisy user-generated text with local distance neighbor feature. *Neurocomputing*, 382:1–11.

Claire Nédellec, Philippe Bessières, Robert R. Bossy, Alain Kotoujansky, and Alain Pierre Manine. 2006. Annotation guidelines for machine learning-based named entity recognition in microbiology. In *Proceeding of Data and Text Mining for Integrative Biology Workshop 17. European Conference on Machine Learning 10. European Conference on Principles and Practice of Knowledge Discovery in Databases*, Workshop on data and text mining for integrative biology, Berlin, Germany. Springer - Verlag.

Peter S. Ostling, Kelly S Davidson, Best O Anyama, Erik M. Helander, M. Q. Wyche, and A. Kaye. 2018. America's opioid epidemic: a comprehensive review and look into the rising crisis. *Current Pain and Headache Reports*, 22:1–7.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Zahra Rezaei, H. Ebrahimpour-Komleh, B. Eslami, Ramyar Chavoshinejad, and M. Totonchi. 2020. Adverse drug reaction detection in social media by deep learning methods. *Cell Journal (Yakhteh)*, 22:319 – 324.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nakatani Shuyo. 2010. Language detection library for Java.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

U.S. Food and Drug Administration. 2021. Human drugs. definition of the U.S. Food and Drug Administration (fda). `https://www.fda.gov/industry/regulated-products/human-drugs`.

A. Viera and J. Garrett. 2005. Understanding interobserver agreement: the Kappa statistic. *Family medicine*, 37 5:360–3.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Dataset statistics

Figure 6 shows a histogram of the number of tokens per document, Figure 7 and Table 8 present the most frequent drug entities. Figure 5 contains the entity frequency histogram.

Entity	Frequency
cocaine	476
thc	435
mdma	428
cannabis	365
indica	350
sativa	336
hash	239
weed	215
lsd	214
hybrid	175
xanax	154
cbd	151
marijuana	150
og kush	120
xtc	116
sour diesel	116
alprazolam	104
heroin	97
oxycodone	96
speed	93
adderall	93

Table 8: Most frequent drug entities in the DreamDrug dataset (lowercased)

B Custom preprocessing

Listing 1: Preprocessing function for the removal of links and special characters from text

```
def remove_unwanted_elements(text):
  ft = re.sub(
    r'https?:\/\/\S*[\r\n]*', '', text)
  ft = re.sub(
    r'\S*.onion\S*[\r\n]*', '', ft)
  ft = re.sub(
    r'[\+!~@#$%^&*()={}\[\]:;<.>?\`"]',
    '', ft)

  ft = re.sub(',', '&#44', ft)

  ft = re.sub(r'[-]+', '-', ft)
  ft = re.sub(r'[_]+', '_', ft)
  return ft
```

Listing 2: Preprocessing functions for replacing email addresses

```
from faker import Faker
# create random name
def get_fake_name():
    # can create fake names
    fake = Faker()
    fake_name = fake.name().replace(
        "␣", "")
    if random.randint(1, 10) % 2 == 0:
        fake_name = fake_name.lower()
    if random.randint(1, 10) % 3 == 0:
        fake_name = "%s%d" % (
            fake_name.lower(),
            random.randint(10, 99))
    return fake_name

def replace_mails(text):
    domain = [
        '.com','.de','.ru','.org']
    vendor = [
        'gmail', 'air', 'wing',
        'microsoft', 'hotmail',
        'outlook']
    mail = get_fake_name() + '@' +
        vendor[random.randint(
            0, len(vendor)-1)] +
        domain[random.randint(
            0, len(domain)-1)]
    return re.sub(
        r'\S+@\S+\s?', mail, text)
```

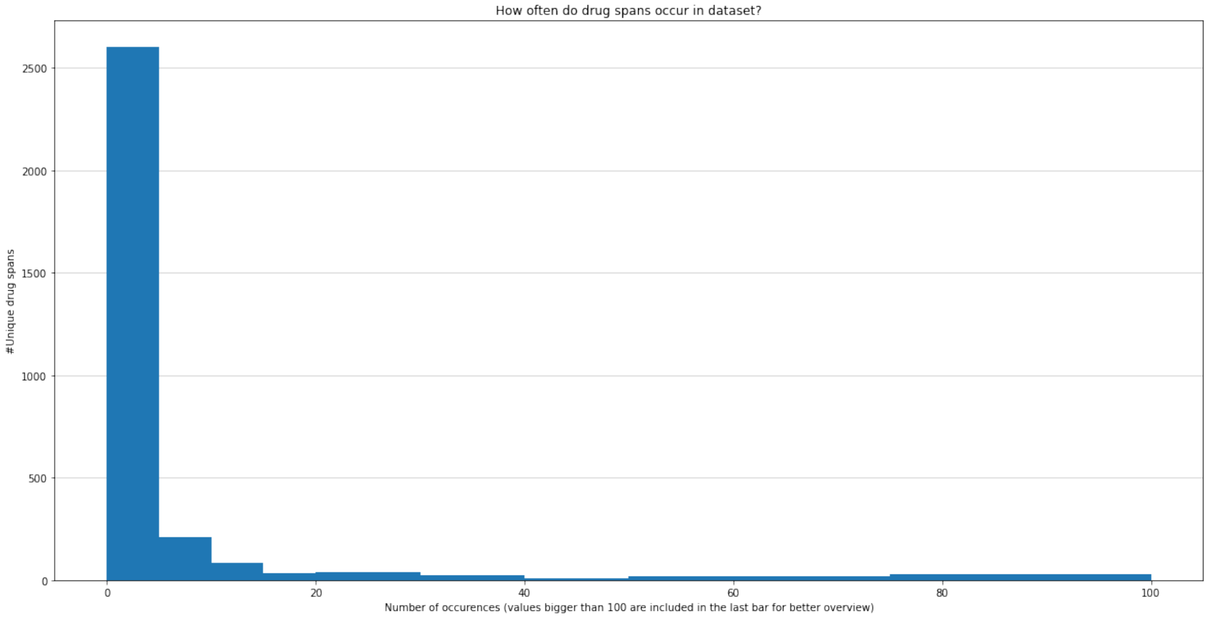


Figure 5: Frequency distribution of drug entities

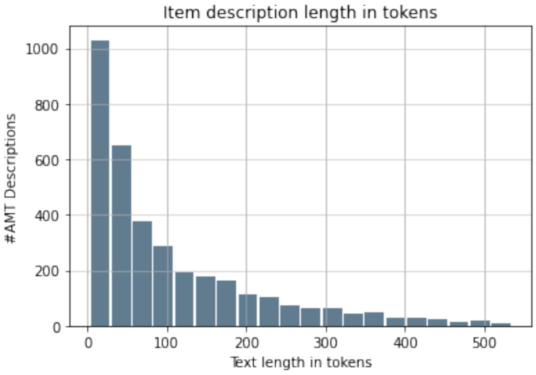


Figure 6: Distribution of document length (number of tokens) in the DreamDrug dataset

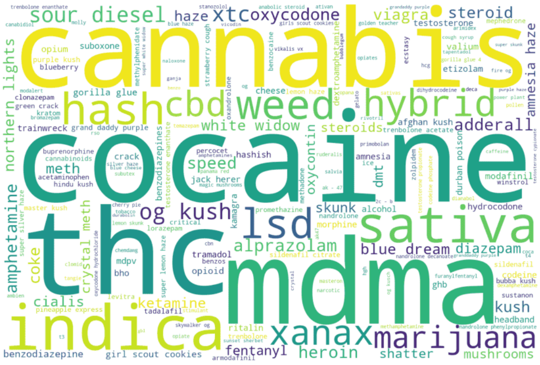


Figure 7: Word-cloud of the most frequent entities in the DreamDrug dataset

C Annotation guidelines

The following pages show the long version of the annotation guidelines as they were presented to the annotators of Appen and to AMT workers clicking on the *Instructions* button on the interface shown in Figure 3 (Section 3.3).

Annotator Guidelines - Appen

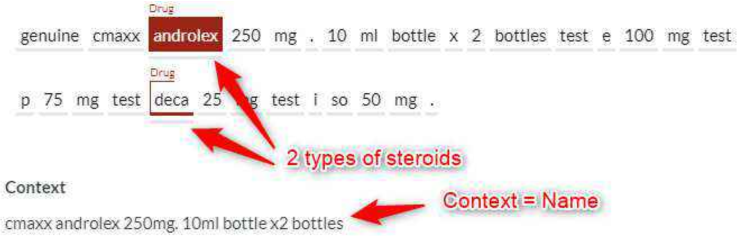
Instructions ▴

Overview

In this task you will use the text annotation tool to highlight drugs in texts from the Darknet. This highlighting means you label pieces of text (called "token") in product descriptions of Darknet Markets. The items have a name and a description. We want you to label all DRUGS you can find in the description.

HOW TO DO THIS?

1. **Read the Context** below the text to get an idea of which items could be contained in the text



Drug

genuine cmaxx androlex 250 mg . 10 ml bottle x 2 bottles test e 100 mg test

Drug

p 75 mg test deca 25 mg test i so 50 mg .

2 types of steroids

Context

cmaxx androlex 250mg. 10ml bottle x2 bottles

Context = Name

2. **Read the provided drug description and label ALL drugs --> Not just the one in the context.**
3. **Are you unsure about what is a drug?** No problem - Click on the info of the "Drug" Class or **re-open this text.**

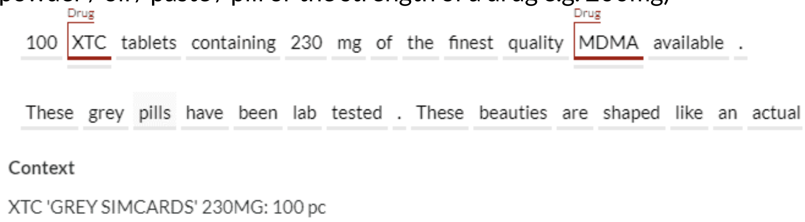
Annotation Guidelines

Label every unique drug name you can find! Typical Drug Examples are:

- Marijuana Strains or types - s e.g. "Gorilla Glue", "Indica / Sativa / Hybrid", "Super silver Haze" or "Girl Scout Cookies"
- Types of Drugs or Agents e.g. "Cocaine", "Crystal Meth", "MDMA" or "THC"
- Legal or partial legal drugs e.g. "Alcohol", "Testosterone Enthalate", "Aspirin" or "Viagra"
- Slang names e.g. Coke, Meth, mushrooms, hash, weed

Label ONLY the drug - Descriptions or Adjectives are not part of the drug

(like "Colour, Form or If its powder / oil / paste / pill or the strength of a drug e.g. 200mg)



Drug Drug

100 XTC tablets containing 230 mg of the finest quality MDMA available .

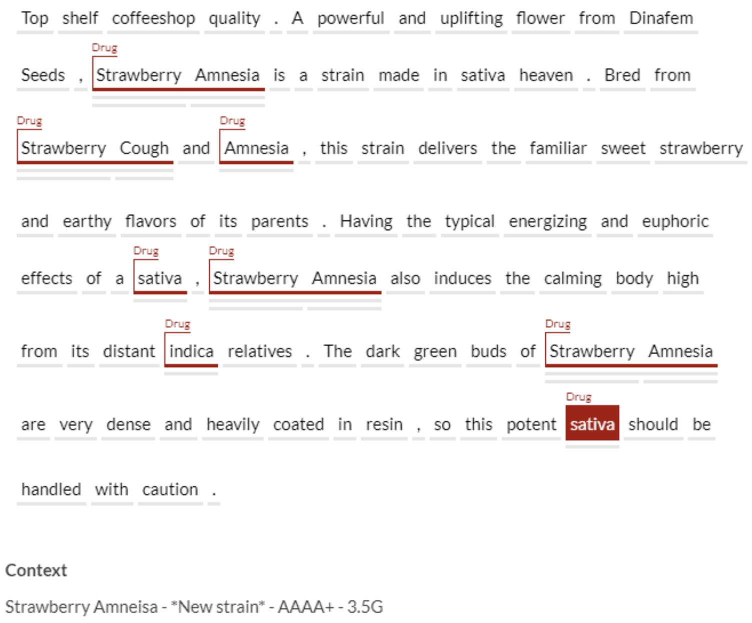
These grey pills have been lab tested . These beauties are shaped like an actual

Context

XTC 'GREY SIMCARDS' 230MG: 100 pc

We want you to label ALL drugs in the text - Pay attention to label drugs with multiple words as one drug. (Strawberry is not drug without amnesia)

If there is a '.' or '-' contained in the middle of a name label it as well. "Strawberry . Amnesia" or " Strawberry - Amnesia" would also be a drug.



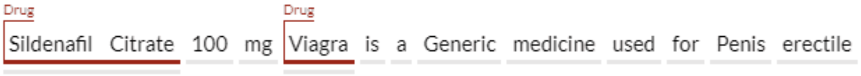
Top shelf coffeeshop quality . A powerful and uplifting flower from Dinafem Seeds , Strawberry Amnesia is a strain made in sativa heaven . Bred from Strawberry Cough and Amnesia , this strain delivers the familiar sweet strawberry and earthy flavors of its parents . Having the typical energizing and euphoric effects of a sativa , Strawberry Amnesia also induces the calming body high from its distant indica relatives . The dark green buds of Strawberry Amnesia are very dense and heavily coated in resin , so this potent sativa should be handled with caution .

Context

Strawberry Amneisa - *New strain* - AAAA+ - 3.5G

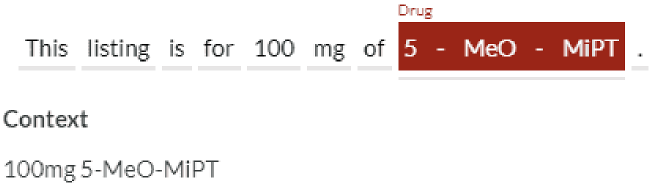
Chemical Details are part of the drug.

- Sildenfanil Citrate is a single drug and need to be labelled together (not separate)



Sildenafil Citrate 100 mg Viagra is a Generic medicine used for Penis erectile

- Drug Chemicals can also be drugs on their own. - AND Include "-" or "." in the labelling

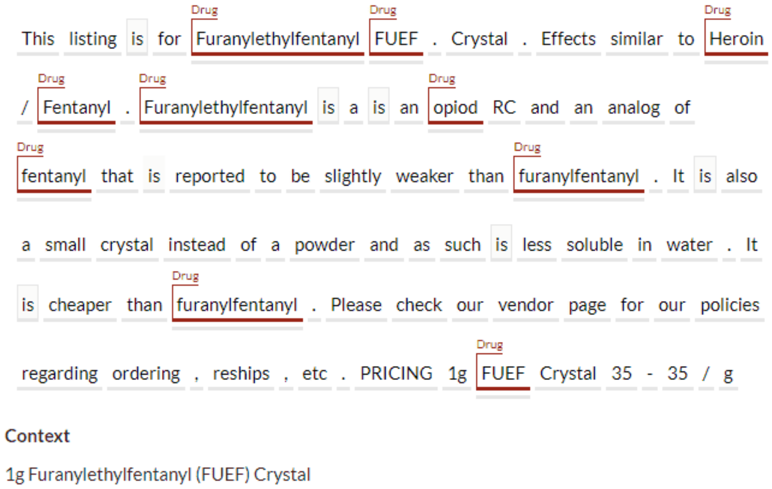


This listing is for 100 mg of 5 - MeO - MiPT .

Context

100mg 5-MeO-MiPT

Abbreviations / Short names of drugs shall be labelled. e.g. XTC, BTH or

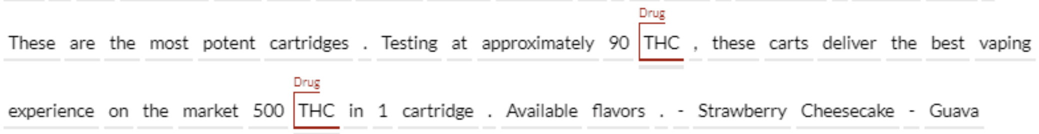


This listing is for Furanylethylfentanyl FUEF . Crystal . Effects similar to Heroin / Fentanyl . Furanylethylfentanyl is a is an opiod RC and an analog of fentanyl that is reported to be slightly weaker than furanylfentanyl . It is also a small crystal instead of a powder and as such is less soluble in water . It is cheaper than furanylfentanyl . Please check our vendor page for our policies regarding ordering , reships , etc . PRICING 1g FUEF Crystal 35 - 35 / g

Context

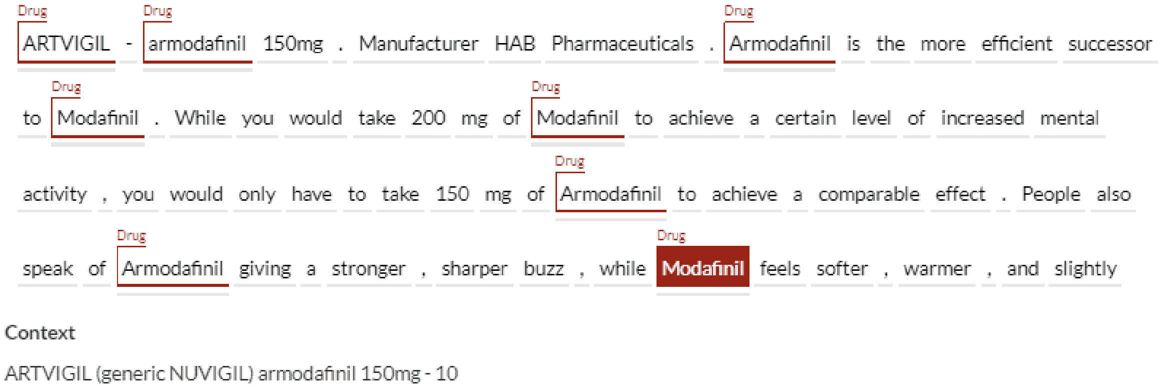
1g Furanylethylfentanyl (FUEF) Crystal

Active Agents (The ingredient which makes you high) shall be labelled as drugs. e.g. CBD, THC, Diazepam or DMT


100 x. 05g cartridges filled with the highest quality distillate oil , 16 each - ORGANIC - Solvent FREE .
These are the most potent cartridges . Testing at approximately 90 THC , these carts deliver the best vaping
experience on the market 500 THC in 1 cartridge . Available flavors . - Strawberry Cheesecake - Guava

Official Marketing names ARE drugs

This includes all kinds of marketing names in different countries or commercial marijuana strain names

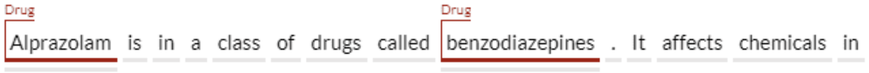

ARTVIGIL - armodafinil 150mg . Manufacturer HAB Pharmaceuticals . Armodafinil is the more efficient successor
to Modafinil . While you would take 200 mg of Modafinil to achieve a certain level of increased mental
activity , you would only have to take 150 mg of Armodafinil to achieve a comparable effect . People also
speak of Armodafinil giving a stronger , sharper buzz , while Modafinil feels softer , warmer , and slightly

Context

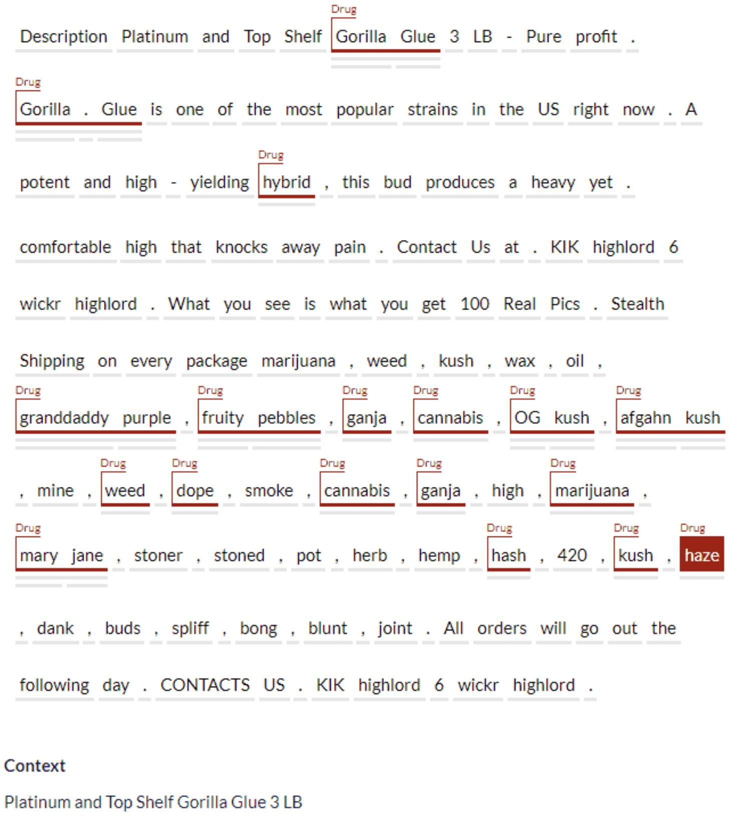
ARTVIGIL (generic NUVIGIL) armodafinil 150mg - 10

Drug Classes are Drugs

- Bezondiazipines, Opiates, Steroids, Indica / Sativa / Hybrid - ARE Drugs


Alprazolam is in a class of drugs called benzodiazepines . It affects chemicals in

Slang names ARE drugs e.g. ganja or dope

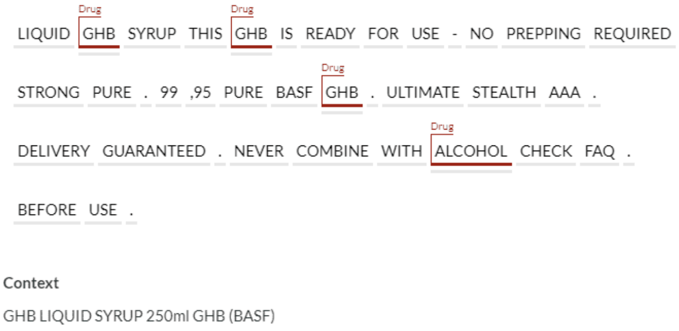

Description Platinum and Top Shelf Gorilla Glue 3 LB - Pure profit .
Gorilla . Glue is one of the most popular strains in the US right now . A
potent and high - yielding hybrid , this bud produces a heavy yet .
comfortable high that knocks away pain . Contact Us at . KIK highlord 6
wickr highlord . What you see is what you get 100 Real Pics . Stealth
Shipping on every package marijuana , weed , kush , wax , oil ,
granddaddy purple , fruity pebbles , ganja , cannabis , OG kush , afgahn kush
, mine , weed , dope , smoke , cannabis , ganja , high , marijuana .
mary jane , stoner , stoned , pot , herb , hemp , hash , 420 , kush , haze
, dank , buds , spliff , bong , blunt , joint . All orders will go out the
following day . CONTACTS US . KIK highlord 6 wickr highlord .

Context

Platinum and Top Shelf Gorilla Glue 3 LB

Legal Drugs ARE Drugs

- In this example its alcohol, but it could also be Caffeine, Aspirin or tabacco.



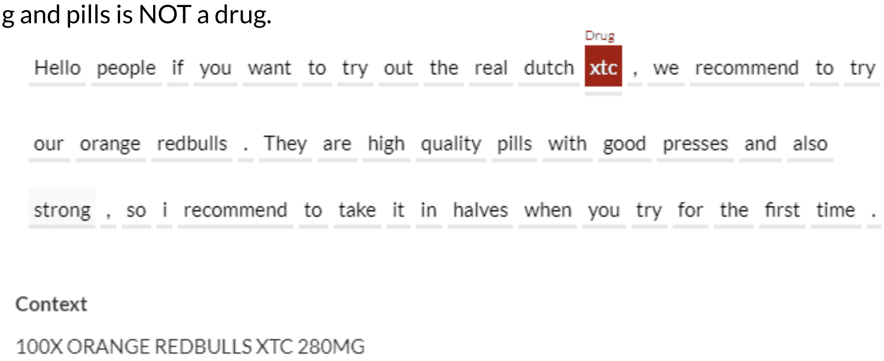
LIQUID GHB SYRUP THIS GHB IS READY FOR USE - NO PREPPING REQUIRED
STRONG PURE . 99 ,95 PURE BASF GHB . ULTIMATE STEALTH AAA .
DELIVERY GUARANTEED . NEVER COMBINE WITH ALCOHOL CHECK FAQ .
BEFORE USE .

Context

GHB LIQUID SYRUP 250ml GHB (BASF)

If a drug is mentioned but doesn't clearly indicate which drug it is, do NOT label it. (e.g. Orange Red bull Pills)

"Fishscale Powder", "These Mitsubishi Pills" or "This strain" will refer to a drug but its not clear, which drug it is exactly. In this case only XTC IS a drug and pills is NOT a drug.



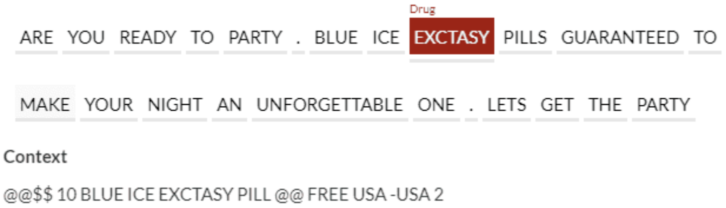
Hello people if you want to try out the real dutch xtc , we recommend to try
our orange redbulls . They are high quality pills with good presses and also
strong , so i recommend to take it in halves when you try for the first time .

Context

100X ORANGE REDBULLS XTC 280MG

The form / strength / adjectives of the drug shall NOT be labelled.

- At Ecstasy(XTC) it doesn't matter if they are Green Mitsubishi with 200mg



ARE YOU READY TO PARTY . BLUE ICE EXCTASY PILLS GUARANTEED TO
MAKE YOUR NIGHT AN UNFORGETTABLE ONE . LETS GET THE PARTY

Context

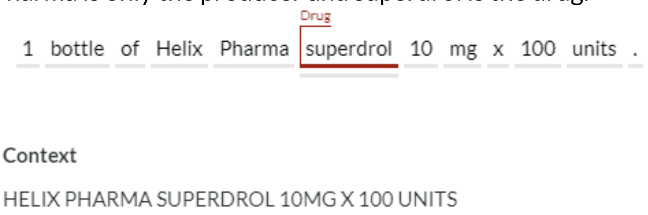
@@$$ 10 BLUE ICE EXCTASY PILL @@ FREE USA -USA 2

- Or that they have the imprint "OT-20" or "V-3605" on them. Those are NOT drugs.

The Producer or Manufacturer of the drug, is NOT a Drug

This also excludes creations from Drugdealers like Drcokes megasniff or MegaBelt IsolatorVapes. None of those are drugs.

In the example below Helix Pharma is only the producer and superdrol is the drug.



1 bottle of Helix Pharma superdrol 10 mg x 100 units .

Context

HELIX PHARMA SUPERDROL 10MG X 100 UNITS

Natural chemicals which are part of your body are NOT a drug.

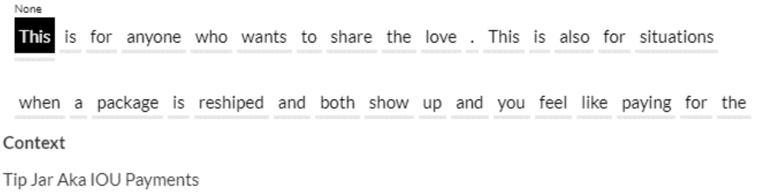
- Dopamin or estrogen do something to your body, but are NOT produced as drug and therefore we DON'T label them.

Processed Items like edibles are not always a drug.

Only the words which clearly refer to a drug are drugs. Brownies, Vapes or Candies are by itself not a drug, only the form!

NO DRUG in the TEXT?? - IMPORTANT

- If you cannot find any drug in the text, just annotate the first word with the tag "NONE" and continue.



This is for anyone who wants to share the love . This is also for situations
when a package is reshiped and both show up and you feel like paying for the

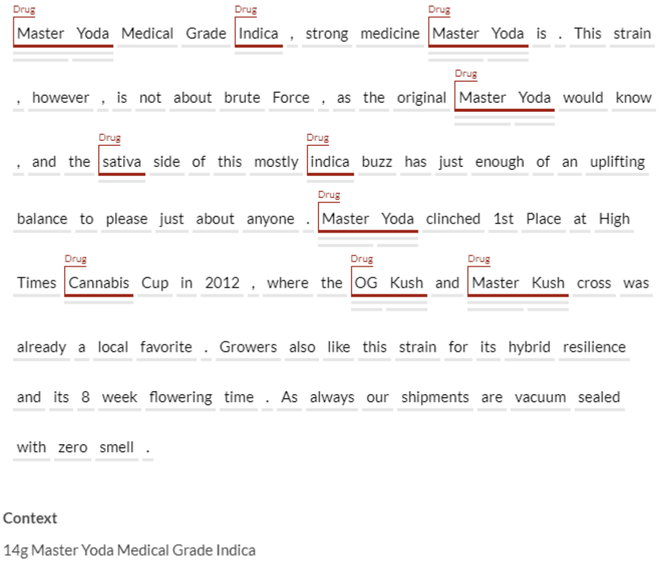
Context

Tip Jar Aka IOU Payments

HOW TO LABEL? - IMPORTANT

We want to label tokens which belong together as ONE single span:

Two tokens which are part of the same drug e.g. the "Master Yoda" Cannabis strain shall be labelled as single span. (Same goes for OG Kush or Master Kush etc.)

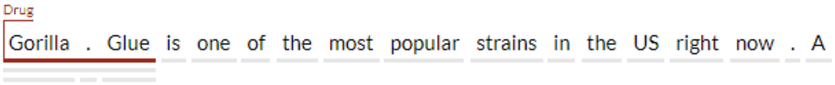


Master Yoda Medical Grade Indica , strong medicine Master Yoda is . This strain
, however , is not about brute Force , as the original Master Yoda would know
, and the sativa side of this mostly indica buzz has just enough of an uplifting
balance to please just about anyone . Master Yoda clinched 1st Place at High
Times Cannabis Cup in 2012 , where the OG Kush and Master Kush cross was
already a local favorite . Growers also like this strain for its hybrid resilience
and its 8 week flowering time . As always our shipments are vacuum sealed
with zero smell .

Context

14g Master Yoda Medical Grade Indica

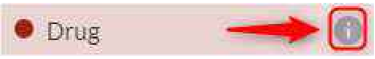
If there is a character (e.g. '.' or '-') inside of drug name -- include it in the labelling. The labels are not allowed to be separated!



Drug

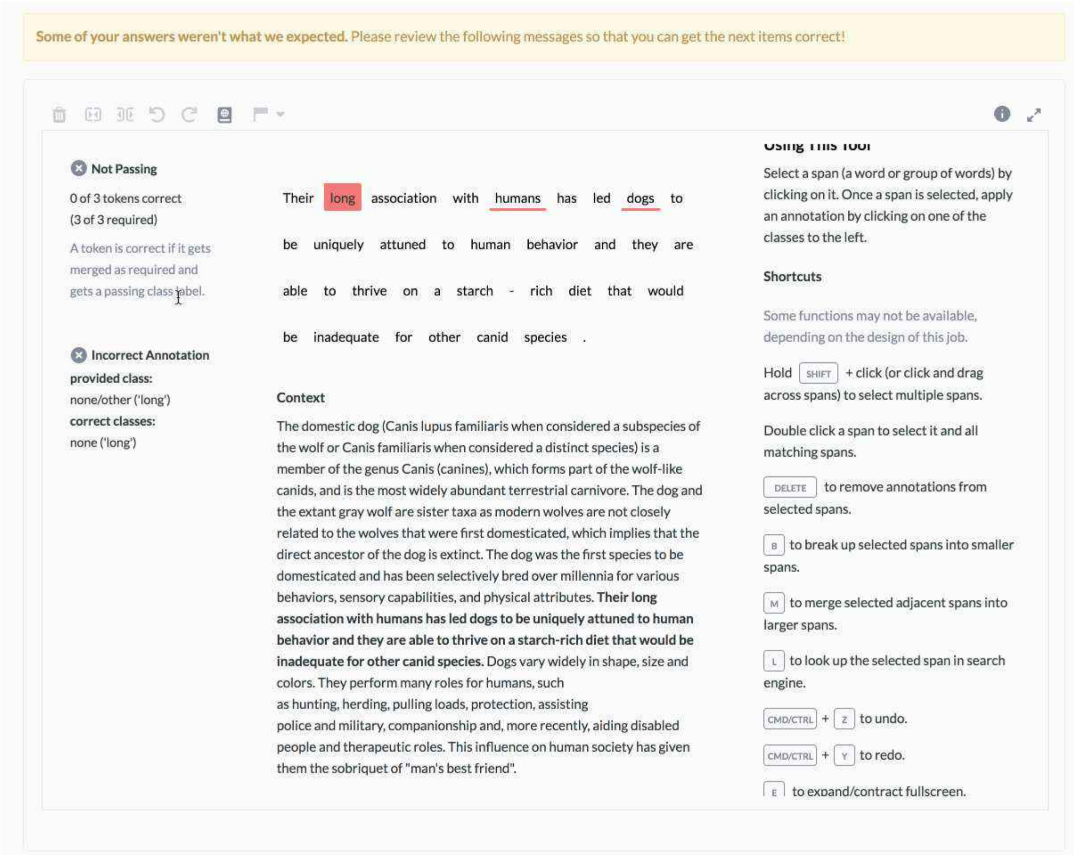
Gorilla . Glue is one of the most popular strains in the US right now . A

Are you unsure about what is a drug? No problem - RE-OPEN thiswholesome descriptionfor details or check the info of the "Drug" Class.



Test Question Review

When you get a test question wrong, you'll be able to review your mistakes in a special version of the tool. The errors you made will be underlined in red.



Some of your answers weren't what we expected. Please review the following messages so that you can get the next items correct!

Not Passing

0 of 3 tokens correct
(3 of 3 required)

A token is correct if it gets merged as required and gets a passing class label.

Incorrect Annotation
provided class:
none/other ('long')
correct classes:
none ('long')

Their long association with humans has led dogs to be uniquely attuned to human behavior and they are able to thrive on a starch - rich diet that would be inadequate for other canid species .

Context

The domestic dog (Canis lupus familiaris when considered a subspecies of the wolf or Canis familiaris when considered a distinct species) is a member of the genus Canis (canines), which forms part of the wolf-like canids, and is the most widely abundant terrestrial carnivore. The dog and the extant gray wolf are sister taxa as modern wolves are not closely related to the wolves that were first domesticated, which implies that the direct ancestor of the dog is extinct. The dog was the first species to be domesticated and has been selectively bred over millennia for various behaviors, sensory capabilities, and physical attributes. **Their long association with humans has led dogs to be uniquely attuned to human behavior and they are able to thrive on a starch-rich diet that would be inadequate for other canid species.** Dogs vary widely in shape, size and colors. They perform many roles for humans, such as hunting, herding, pulling loads, protection, assisting police and military, companionship and, more recently, aiding disabled people and therapeutic roles. This influence on human society has given them the sobriquet of "man's best friend".

Using This Tool

Select a span (a word or group of words) by clicking on it. Once a span is selected, apply an annotation by clicking on one of the classes to the left.

Shortcuts

Some functions may not be available, depending on the design of this job.

Hold SHIFT + click (or click and drag across spans) to select multiple spans.

Double click a span to select it and all matching spans.

DELETE to remove annotations from selected spans.

B to break up selected spans into smaller spans.

M to merge selected adjacent spans into larger spans.

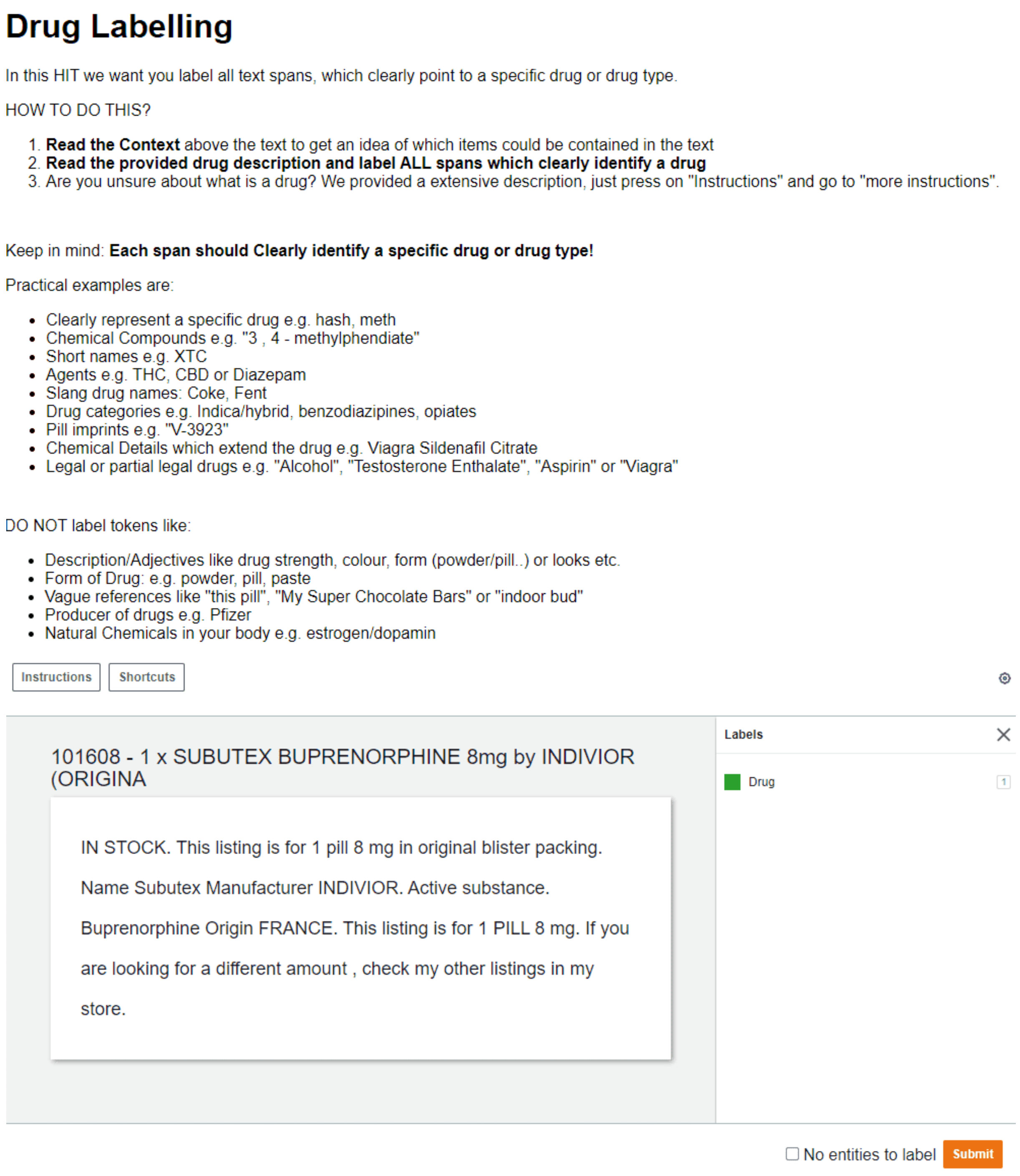
L to look up the selected span in search engine.

CMD/CTRL + Z to undo.

CMD/CTRL + Y to redo.

E to expand/contract fullscreen.

Annotator Guidelines Amazon Mechanical Turk



Drug Labelling

In this HIT we want you label all text spans, which clearly point to a specific drug or drug type.

HOW TO DO THIS?

1. **Read the Context** above the text to get an idea of which items could be contained in the text
2. **Read the provided drug description and label ALL spans which clearly identify a drug**
3. Are you unsure about what is a drug? We provided a extensive description, just press on "Instructions" and go to "more instructions".

Keep in mind: **Each span should Clearly identify a specific drug or drug type!**

Practical examples are:

- Clearly represent a specific drug e.g. hash, meth
- Chemical Compounds e.g. "3 , 4 - methylphendiate"
- Short names e.g. XTC
- Agents e.g. THC, CBD or Diazepam
- Slang drug names: Coke, Fent
- Drug categories e.g. Indica/hybrid, benzodiazipines, opiates
- Pill imprints e.g. "V-3923"
- Chemical Details which extend the drug e.g. Viagra Sildenafil Citrate
- Legal or partial legal drugs e.g. "Alcohol", "Testosterone Enthalate", "Aspirin" or "Viagra"

DO NOT label tokens like:

- Description/Adjectives like drug strength, colour, form (powder/pill..) or looks etc.
- Form of Drug: e.g. powder, pill, paste
- Vague references like "this pill", "My Super Chocolate Bars" or "indoor bud"
- Producer of drugs e.g. Pfizer
- Natural Chemicals in your body e.g. estrogen/dopamin

101608 - 1 x SUBUTEX BUPRENORPHINE 8mg by INDIVIOR (ORIGINA

IN STOCK. This listing is for 1 pill 8 mg in original blister packing. Name Subutex Manufacturer INDIVIOR. Active substance. Buprenorphine Origin FRANCE. This listing is for 1 PILL 8 mg. If you are looking for a different amount , check my other listings in my store.

Note: The long version of the annoation guidelines was accessible over the „Instructions“ button. The long version was equal to the long Annotation guidelines from Appen.