Covering Cracks in Content Moderation: Delexicalized Distant Supervision for Illicit Drug Jargon Detection

Minkyoo Song
Korea Advanced Institute of Science & Technology
Daejeon, Republic of Korea
minkyoo9@kaist.ac.kr

Eugene Jang
S2W Inc.
Seongnam, Republic of Korea
genesith@s2w.inc

Jaehan Kim
Korea Advanced Institute of Science & Technology
Daejeon, Republic of Korea
jaehan@kaist.ac.kr

Seungwon Shin
Korea Advanced Institute of Science & Technology
Daejeon, Republic of Korea
claude@kaist.ac.kr

Abstract

In light of rising drug-related concerns and the increasing role of social media, sales and discussions of illicit drugs have become commonplace online. Social media platforms hosting user-generated content must therefore perform content moderation, which is a difficult task due to the vast amount of jargon used in drug discussions. Previous works on drug jargon detection were limited to extracting a list of terms, but these approaches have fundamental problems in practical application. First, they are trivially evaded using word substitutions. Second, they cannot distinguish whether euphemistic terms (*pot*, *crack*) are being used as drugs or as their benign meanings. We argue that drug content moderation should be done using contexts, rather than relying on a banlist. However, manually annotated datasets for training such a task are not only expensive but also prone to becoming obsolete. We present JEDIS, a framework for detecting illicit drug jargon terms by analyzing their contexts. JEDIS utilizes a novel approach that combines distant supervision and delexicalization, which allows JEDIS to be trained without human-labeled data while being robust to new terms and euphemisms. Experiments on two manually annotated datasets show JEDIS significantly outperforms state-of-the-art word-based baselines in terms of F1-score and detection coverage in drug jargon detection. We also conduct qualitative analysis that demonstrates JEDIS is robust against pitfalls faced by existing approaches.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → **Information systems applications**.

Keywords

Drug Jargon Detection; Content Moderation; Delexicalized Distant Supervision; Natural Language Processing

ACM Reference Format:
Minkyoo Song, Eugene Jang, Jaehan Kim, and Seungwon Shin. 2025. Covering Cracks in Content Moderation: Delexicalized Distant Supervision for Illicit Drug Jargon Detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3690624.3709183

KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1245-6/25/08
https://doi.org/10.1145/3690624.3709183

1 Introduction

Drug-related deaths are at record highs, with annual drug overdose deaths in the US more than doubling since 2015 [27, 5]. In March 2022, the United Nations urged governments to further regulate social media, citing a link between exposure to social media and drug abuse [26]. Illicit sales and discussion of drugs have become prevalent online not only on anonymous marketplaces [41, 2] but also on social media [30, 54]. On top of illegal drugs, online discussion of legal drugs for non-medical recreational usage has also become commonplace [42, 35]. Without sufficient moderation, social media can expose the general public to sales of illegal drugs [17, 18] and dangerous drug misinformation [3].

Content moderation is a difficult and labor-intensive task. Although content moderators often face psychological distress from large volumes of explicit content [43], automating moderation of user-generated content is challenging. Even for human moderators, understanding the slang and jargon used in illegal content requires significant domain knowledge. Previous works have addressed this by discovering jargon terms [46, 48, 52], interpreting jargon terms [37, 38], or both [54]. Practically, however, moderation systems relying on a banlist of jargon terms suffer from two main limitations. Figure 1 shows an example input that demonstrates the limitations of a word-based moderation system. First, word-based restrictions are easily bypassed by alternative terms for moderated words [19]. This not only includes newly emerging words, but also misspellings of known banned words (*cokaine*, *c0caine*). Second, euphemistic jargon has benign usages that can cause false positives. While *speed* requires moderation when referring to amphetamines, it should not be moderated when referring to quickness. In order to effectively address such limitations, moderation systems must be trained to evaluate words within their specific context. Thus, for more practical content moderation automation, we define *jargon detection* as a sentence-level task of identifying if a word, within its specific context, is used as drug jargon. To the best of our knowledge, this sentence-level approach for deriving contextual information is a novel direction in drug content moderation research.

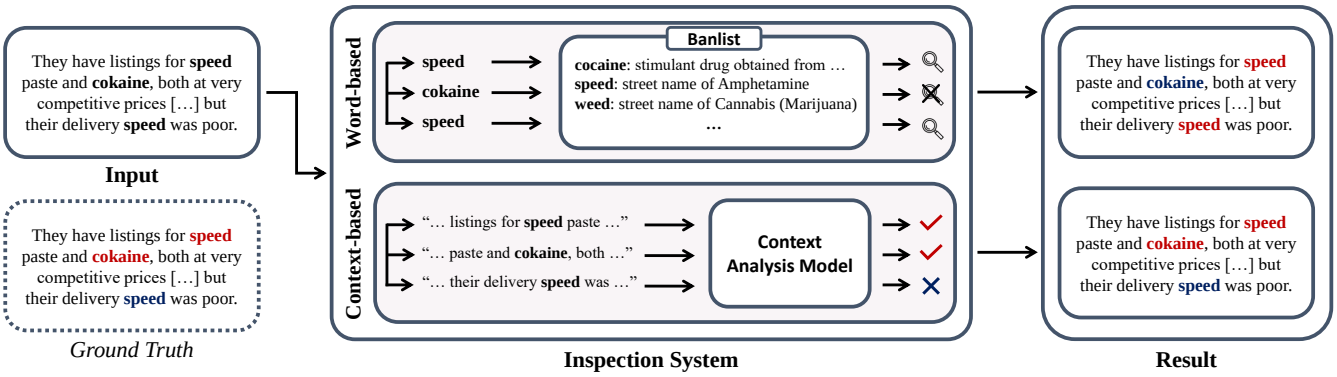


Figure 1: Content moderation procedures of word-based system and context-based system. Bolded words in *Input* were predicted by each system. In the *Result*, words detected as drug jargon are colored red, and words predicted as non-drugs are colored blue.

We propose **JEDIS** (Jargon and Euphemism Detection of Illicit Substances), a novel drug jargon detection framework that performs context-aware detection of jargon instances. JEDIS utilizes distant supervision to keep up with complex and fast-evolving drug slang without depending on expensive manual annotations from domain experts. Drug jargon contexts from an unlabeled text corpus are learned by leveraging only a small list of known drug seed terms. Given an incomplete and small knowledge base for distant supervision, the model may overfit to the seed terms. To generalize detection, JEDIS employs an innovative training strategy that effectively utilizes learned contexts. Specifically, JEDIS uses two independent modules to represent contextual and lexical information separately. The delexicalized context-only module mitigates the excessive focus on the target word by extracting context information alone. The word attribute module preserves target word information to address imprecise detection caused by delexicalization. This allows JEDIS to robustly detect unrecognized terms and accurately distinguish between euphemisms.

We construct two human-labeled datasets to evaluate moderation approaches on user-generated drug context sentences. Our findings suggest that word-based methods perform poorly with content in the wild due to their limitations. We find that JEDIS substantially outperforms state-of-the-art moderation approaches in F1-score across both evaluation datasets. Specifically, JEDIS successfully detects an average of 4.23 times more various drug jargon terms compared to word-based baselines. It also highlights the need for a tailored context-based method for effective drug jargon detection. Qualitative analysis shows JEDIS detects jargon in scenarios that previous approaches struggled with.

Our contributions are as follows:

- Propose JEDIS, a drug jargon detection framework robust to unseen words and euphemisms by effectively combining contextual and lexical information without requiring labeled training data.
- Propose delexicalized distant supervision for JEDIS training, enabling generalization to new terms without overfitting to a small set of drug seed terms.
- Create two datasets of forum texts annotated for drug jargon instances to evaluate drug jargon detection[1].
- Empirically demonstrate limitations of word-based moderation in practical scenarios, which suggest the necessity of context-based moderation.
- Empirically show shortcomings of recent language model with prompts in drug jargon detection, underscoring the need for a tailored detection method.

2 Background

2.1 Preliminaries

We define **drug jargon** as words used to refer to illicit drug substances or abused legal drug substances. This can include both explicit substance names (*MDMA*, *ketamine*) and street names (*Molly*, *ket*). Some jargon terms are **euphemisms** (*speed*, *pot*), polysemous words that have alternate innocuous meanings. It should be noted that slang terms are not necessarily euphemisms (*fent* is slang for *fentanyl* but has no benign alternative usage). We formalize the task of **drug jargon detection** as a binary classification problem. Given a sentence and a word in the sentence, determine whether the word is used as a drug jargon in the specific context. This task is contrary to past works, which were typically focused on extracting a list of drug jargon terms from a target corpus.

2.2 Related Work

Illegal content in social media. Detecting abusive content on social media is a challenging task [40], especially when it is implicitly abusive [45, 34]. Works on hate speech have identified the usage of euphemisms [19, 20] and emojis [14] to circumvent moderation. Previous works on Twitter data found illegal trafficking of fentanyl [17] and opioids [18]. Previous research on Instagram data identified drug traffickers using text [16], multimodal combinations of text and image [10], and heterogeneous graphs that incorporate meta-knowledge [30]. These works are limited to finding drug traffickers, and not applicable to drug discussions on social media. Such discussions are prone to inaccurate and biased information that can lead to fatal results [3].

[1]Our code is available at https://github.com/minkyoo9/JEDIS.

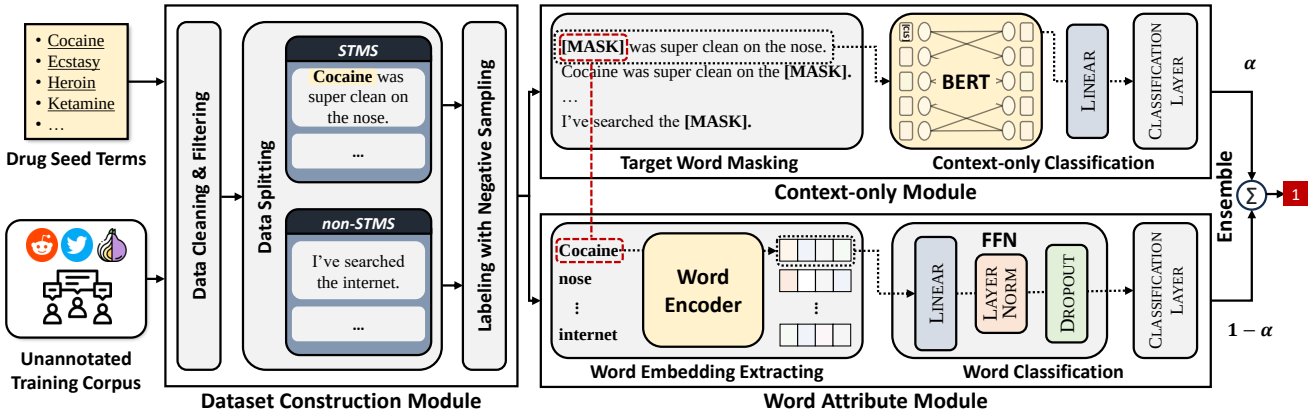


Figure 2: Overview of JEDIS.

Jargon detection. The languages used by underground communities and marketplaces have been of significant interest to researchers [8, 12, 11]. A common line of jargon detection research follows a set expansion formulation: expanding a set of known drug names by finding similar entities. Euphemisms of known drugs were found using masked language modeling (MLM) on informative sentences [53, 54]. Other works have used a set of known jargon to find new jargon from darknet websites [13], underground market groups [52], and search engine optimization sites [46]. Another line of approach is to compare text corpora to identify atypical word usage. Yuan et al. [48] leveraged differences in Word2Vec embeddings between different types of corpora to discover words that are used in semantically different ways. Other works used word probabilities to translate "dark" jargon to their "clean" counterparts across corpora [37, 38]. However, all these methods primarily generate a list of drug jargon after analyzing a target corpus. We contend that such word-based approaches are limited in practical application and advocate for a context-based moderation strategy.

Distant supervision and delexicalization. Large language models (LLMs) [4] have greatly progressed natural language processing, but such models still require annotated datasets to approach complex tasks such as content moderation. However, annotated datasets are expensive and prone to become obsolete as the language of the domain changes. Distant supervision solves the problem of annotation by automatically labeling data using an existing knowledge base. Distant supervision has been utilized for relation extraction [25, 31, 32] and named entity recognition [47, 22].

A known challenge in distant supervision is that using an incomplete knowledge base generates noisy labels [24]. Our problem setting is an extreme case where the knowledge base of drug names is very small and incomplete, unlike the Freebase database used for relation extraction containing 116 million relation instances [25], or the administrative privileges of the Baidu search engine used for jargon detection [46]. Such extensive resources are often infeasible for moderators. To generalize detection to unknown terms instead of simply memorizing known words, we propose a novel approach that integrates delexicalization into the distantly supervised data. Delexicalization is the concept of replacing the target entities from a text, so that tasks focus on contexts without relying on the target entities. This technique has been used to improve fact verification [44] and spoken language understanding [33].

3 JEDIS

We present an overview of JEDIS in Figure 2. In *dataset construction module*, an unannotated training corpus of drug domain texts is cleaned and filtered to maintain the quality of the dataset. The corpus is then split and carefully labeled using drug seed terms and negative sampling to construct the training dataset. Then, JEDIS uses the dataset to train two modules: the *context-only module* and the *word attribute module*. In context-only module, inspected target words are masked from training sentences. These masked sentences are then used to train a BERT-based sequence classification model, without exposing the model to the original target word (i.e., delexicalization). This module is intended to detect drug jargon from contexts alone. On the other hand, the word attribute module utilizes a vector representation of the target word within its context. This is intended to utilize lexical information to understand the semantic role that drug seed terms play in drug-related contexts. To leverage information from both modules, we use an ensemble approach combining both outputs to make final predictions.

3.1 Dataset Construction Module

Since we propose a distant supervision approach for JEDIS training, dataset construction does not require annotated data. Instead, it automatically labels unannotated training data using a knowledge base of drug seed terms, enabling JEDIS to learn drug contexts. To do this, dataset construction module consists of four steps: data cleaning, filtering, splitting, and labeling with negative sampling. In the **Data Cleaning**, we first utilize the sentence tokenizer from NLTK to separate the training corpus into sentences. Then, we remove non-ASCII characters, HTML patterns, and words longer than 16 characters to eliminate noise in the text. In the **Data Filtering**, we filter out sentences longer than 128 tokens when tokenized by the BERT tokenizer. We do this because most sentences over 128 tokens are from spam posts without punctuation. We also filter out sentences shorter than six words, since they have insufficient context. The number of sentences after filtering in our training corpora can be referred to in Section 4.1.

Data Splitting. After cleaning and filtering the training corpus, we use the list of drug seed terms to split it into two pools of sentences: *Seed Term Mentioning Sentences (STMS)* and *sentences without seed terms (non-STMS)*. The former refers to the sentences that mention at least one drug seed term, and the latter refers to the sentences that do not mention any drug seed terms.

Labeling with Negative Sampling. To create *positive* training data, we can simply annotate the seed terms in the STMS. For example, in the STMS *"**Cocaine** was super clean on the nose and gave a great feeling."*, we mark the seed term **cocaine** as a positive sample. On the other hand, creating *negative* training data requires more careful sampling. We sample negative data from both the STMS and non-STMS sentence pools. Neg_{STMS} are negative data created by selecting an alphanumeric noun (that is not a seed term) from STMS. For example, from the previous STMS example sentence, the word **nose** can be chosen to create the sample *"Cocaine was super clean on the **nose** and gave a great feeling."*. Training with Neg_{STMS} teaches the model to precisely locate which words in a drug-related sentence are drug jargon. $Neg_{nonSTMS}$ are negative data that represent contexts without known drug terms. From non-STMS sentences, any alphanumeric noun (like **internet**) can be chosen to create negative samples such as *"I've searched the **internet** and can't find anything too helpful."*.

Since the list of known drug terms is highly incomplete, negative labels are potentially noisy. Empirically, we find that controlling the ratio of negative training data is important to maximize performance in jargon detection. Therefore, we use two hyperparameters to control negative sampling: $\mathcal{R}_{STMS}$ and $\mathcal{R}_{nonSTMS}$. $\mathcal{R}_{STMS}$ is the maximum number of negative samples extracted from a single STMS sentence (if the sentence contains less than $\mathcal{R}_{STMS}$ alphanumeric nouns, all are used). $\mathcal{R}_{nonSTMS}$ represents the ratio of non-STMS sentences selected per STMS sentence. For each selected non-STMS sentence, a single negative sample is made by choosing a random alphanumeric noun as the target. The impact of the two types of negative training data can be regulated by the two hyperparameters, which we can optimize for performance. Further discussion about the negative ratios is included in Section 5 and 6.2.

3.2 Context-only Module

Target Word Masking. Because we construct our training dataset by explicitly utilizing the drug seed terms, directly training on this data can cause a classification model to heavily overfit to the seed terms by simply memorizing them. This is especially the case since our seed term list is relatively small (46) compared to other knowledge bases used for distant supervision. To minimize the effect of seed terms, we *delexicalize* the prediction process by masking the target word (annotated before) with the [MASK] token. Therefore, the classification model must predict whether a target word is drug jargon or not without actually seeing the word in question. This setting removes the lexical influences of the inspected words while emphasizing the contexts.

Context-only Classification. The classifier takes masked sentences and determines if the masked word from a sentence is a drug jargon. To do this, we construct a sequence classification model using BERT [4]. First, the BERT (*bert-base-uncased*) model is adapted to the target domain through a pretraining process. This has been shown to improve performance, especially on domains where the linguistic characteristics deviate from conventional contexts [7].

Then, sentences from the constructed dataset are used to train the model. Each masked training sentence is surrounded by the [CLS] token and the [SEP] token and inputted into the BERT model. The BERT pooled output $\mathbf{o}_{pooled}$ is extracted by passing the [CLS] token embedding $\mathbf{h}_{[CLS]}$ from the last hidden layer of BERT through the linear layer (weights $\mathbf{W}_1 \in \mathbb{R}^{dim \times dim}$ and bias $\mathbf{b}_1 \in \mathbb{R}^{dim}$, dim represents the dimension of BERT embedding) then applying *tanh*.

$$\mathbf{o}_{pooled} = tanh(\mathbf{W}_1 \mathbf{h}_{[CLS]} + \mathbf{b}_1) \tag{1}$$

It represents the context of the whole sentence including the [MASK] token. Finally, we apply classification layer $\mathbf{W}_2 \in \mathbb{R}^{dim \times 1}$ and *sigmoid* on the pooled output to conduct a binary classification task by calculating the probability that the masked word is drug jargon.

$$p_c = sigmoid(\mathbf{W}_2 \mathbf{o}_{pooled}) \tag{2}$$

3.3 Word Attribute Module

Word Embedding Extracting. Relying only on context without seeing the target word may cause false positives[2]. To also utilize lexical information and mitigate false positives, we utilize a word encoder to obtain embeddings of the target word from each sentence. Crucially, the encoder must generate *contextual embeddings*, creating the appropriate representations depending on word context. Thus, we utilize a BERT model pretrained on the target domain as the word encoder. This allows us to utilize the context-sensitive intrinsic information of target words to make jargon detection more precise. This word encoder operates independently from the BERT of the context-only module. The model weights are frozen to prevent overfitting caused by seeing the complete sentences through the entire framework.

Word Classification. We train a feed-forward network and a classification layer that use the dynamic embeddings of the target word. Each word embedding $\mathbf{e}_{word}$ is passed through a feed-forward network consisting of a linear layer (weights $\mathbf{W}_3 \in \mathbb{R}^{dim \times dim}$ and bias $\mathbf{b}_3 \in \mathbb{R}^{dim}$), $LayerNorm$, $tanh$, and $Dropout$.

$$\mathbf{z}_{word} = \mathbf{W}_3 \mathbf{e}_{word} + \mathbf{b}_3 \tag{3}$$

$$\mathbf{o}_{word} = Dropout(tanh(LayerNorm(\mathbf{z}_{word}))) \tag{4}$$

By doing this, information of the target word within its context is represented as $\mathbf{o}_{word}$. This enhanced representation is then used as the input for the classification layer $\mathbf{W}_4 \in \mathbb{R}^{dim \times 1}$ and *sigmoid* function to obtain the probability that the target word is drug jargon.

$$p_w = sigmoid(\mathbf{W}_4 \mathbf{o}_{word}) \tag{5}$$

3.4 Ensemble Training and Prediction

The two probabilities obtained from the context-only and word-attribute modules are used to generate the final prediction of the jargon detection task. In the training phase, probability p_c is used to calculate the loss for the context-only classification model by evaluating the binary cross-entropy loss between the predicted

[2]We further discuss the cases of false positives associated with linguistic challenges in Section 6.1.

Table 1: Summary of data statistics.

	Reddit Drug	Silk Road Forum
Posts	1,271,907	1,227,847
Sentences	6,793,530	3,952,297
Sentences (filtered)	4,799,512	3,057,758
STMS (%)	281,104 (5.86%)	159,944 (5.23%)
Collection period	Feb 2008 - Dec 2017	June 2011 - Oct 2013

probability p_c for the training sample s in training dataset $\mathcal{D}$ and the ground truth label $y \in \{0, 1\}$.

$$\mathcal{L}_c = -\frac{1}{|\mathcal{D}|} \sum_{(s_i, y_i) \in \mathcal{D}} y_i \, log \, p_{c,s_i} + (1 - y_i) \, log \, (1 - p_{c,s_i}) \quad (6)$$

Similarly, we compute the loss for the word classification model by assessing the binary cross-entropy loss between the predicted probability p_w for the training sample and ground truth label.

$$\mathcal{L}_w = -\frac{1}{|\mathcal{D}|} \sum_{(s_i, y_i) \in \mathcal{D}} y_i \, log \, p_{w,s_i} + (1 - y_i) \, log \, (1 - p_{w,s_i}) \quad (7)$$

Then, we incorporate a learnable parameter, α, to compute the final loss as:

$$\mathcal{L} = \alpha * \mathcal{L}_c + (1 - \alpha) * \mathcal{L}_w \quad (8)$$

The classification models within JEDIS are trained by minimizing this ensemble loss. Thus, JEDIS can leverage both contextual and word information with appropriate weights. In the prediction phase, we determine the final probability of the target word being drug jargon by weighting the two probabilities, p_c and p_w, using the previously described parameter α.

$$p = \alpha * p_c + (1 - \alpha) * p_w \quad (9)$$

4 Datasets and Annotation

4.1 Datasets

Drug seed terms. Since we are using the list of drug seed terms for distant supervision, we wish to find explicit drug terms that are unambiguously used as drugs. We refer to the drug slang report published by the Drug Enforcement Administration (DEA) [1], and take only the explicit drug names to create a list of 46 explicit drug terms[3]. The generated list is available in our repository.

Drug forum corpora. We use text data from drug-related forums to train JEDIS on user-generated drug discussions. For effective detection, it's essential that our data includes both drug and non-drug contexts. To assess JEDIS's versatility across settings, we employ two different drug-related forum corpora.

The *Reddit Drug corpus*, provided by Zhu et al.[54], is sourced from the public Reddit archive[9] on Google BigQuery[4]. It comprises posts from 46 subreddits related to drugs and darknet markets, such as 'Drugs', 'DarkNetMarkets', and 'SilkRoad', many of which are now inaccessible due to Reddit's policy change [6]. Some subreddits, like 'Bitcoin' and 'TOR', are only tangentially drug-related but are included to ensure a variety of non-drug contexts. This corpus represents surface web drug discussions.

[3]Illegality may vary depending on local legislation.
[4]https://console.cloud.google.com/bigquery?project=fh-bigquery

Table 2: Statistics of the annotation results. *Only alphanumeric nouns are considered.

	Reddit Drug	Silk Road Forum
Sentences	1,000	1,000
Cohen's Kappa	0.86	0.87
Nouns*	4,559	5,531
Drug jargon (%)	447 (9.8%)	473 (8.6%)
Unique drug jargon	178	169

The *Silk Road Forum corpus*, derived from the darknet market archive by Branwen et al. [2], represents discussions from the Silk Road forum. Silk Road, once the largest drug-related darknet market, was shut down by the FBI in 2013 [29]. Its associated forum hosted categories ranging from 'Legal' and 'Drug safety' to broader subjects like 'Security' and 'Philosophy'. We incorporate posts from all these categories to ensure a mix of drug and non-drug contexts, representing dark web drug discussions. The statistics of our corpora are summarized in Table 1.

4.2 Annotation

To evaluate JEDIS quantitatively, we constructed evaluation datasets by annotating 1,100 randomly selected sentences from each forum corpus, excluding sentences with drug seed terms. Two annotators from a cybersecurity company specializing in monitoring malicious online activities were recruited to annotate the sentences for drug jargon. For quality assurance, we conducted a pilot annotation on the first 100 sentences from each corpus, subsequently constructing annotation guidelines. These pilot sentences were excluded from the final evaluation datasets.

Through this manual annotation, we generated two evaluation datasets, each with 1,000 sentences. Table 2 provides detailed statistics of the annotation results. The inter-annotator agreement registered Cohen's Kappa values of 0.86 and 0.87 for the respective datasets, indicating almost perfect agreement (> 0.8). Both annotation guidelines and the ethical considerations for the dataset construction can be found in Appendix A and D, respectively. Additionally, in Appendix E, we discuss how to address potential subjectivity and bias in the annotation process.

5 Evaluation

5.1 Experimental Setup

Training JEDIS. For each corpus, we create a training dataset using empirically selected hyperparameters $\mathcal{R}_{STMS} = 5$ and $\mathcal{R}_{nonSTMS} = 2$[5]. We also limit the range of learnable parameter α as [0.9, 0.99] to emphasize the delexicalized context-only module[6]. Specifically, we train JEDIS using the *unlabeled* Reddit Drug corpus to evaluate the Reddit Drug annotated dataset. The training data included *329,573* positive instances from STMS, *973,447* negative instances from STMS (Neg_{STMS}), and *659,146* negative instances from non-STMS ($Neg_{nonSTMS}$). For evaluation on the Silk Road Forum annotated dataset, we train the classifier with the *unlabeled* Silk Road Forum corpus. The training data included *193,167* positive instances,

[5]Detailed discussion of these hyperparameters will be in Section 5.2 and 6.2.
[6]We empirically find that a low value of α leads to **significant overfitting**, resulting a low recall.

Table 3: Evaluation results of drug jargon detection task. Boldface represents the best performance and underline represents the second-best performance. The *# Jargon* column represents the number of *unique* jargon terms detected by each approach.

	Reddit Drug				Silk Road Forum			
	Precision	Recall	F1-score	# Jargon	Precision	Recall	F1-score	# Jargon
Word2Vec [23]	**0.8908**	0.2371	0.3746	32	**0.7908**	0.3277	0.4634	39
CantReader [48]	0.7778	0.1879	0.3027	17	0.7208	0.2347	0.3541	22
Zhu et al. [54]	0.5897	0.2573	0.3583	17	0.5026	0.4123	0.4530	27
PETD [15]	0.4760	0.2215	0.3023	28	0.5789	0.2326	0.3318	26
MLM (*w/o pretrain*)	0.5094	0.3043	0.3810	61	0.4569	0.2579	0.3297	57
MLM	0.5241	0.6085	0.5631	98	0.5783	0.6321	0.6040	102
DarkBERT [11]	0.5460	0.5705	0.5580	92	0.4790	0.5793	0.5244	109
GPT4o-mini [28]	0.4307	**0.7919**	0.5579	**129**	0.3920	**0.7907**	0.5242	**137**
JEDIS (*w/o* Neg_{STMS})	0.6052	0.6309	0.6177	105	0.5367	0.6808	0.6002	112
JEDIS (*w/o pretrain&word*)	0.6318	0.5951	0.6129	97	0.5368	0.6321	0.5806	104
JEDIS (*w/o word*)	0.6454	0.6107	0.6276	97	0.5507	0.6998	0.6164	116
JEDIS (*w/o pretrain*)	0.6475	0.5794	0.6116	92	0.5573	0.6068	0.5810	101
JEDIS	0.6659	0.6197	**0.6419**	101	0.5805	0.6934	**0.6320**	113

615,996 Neg_{STMS}, and *386,334* $Neg_{nonSTMS}$. For early stopping, we split the data further into train/valid sets with an 8:2 ratio. More detailed experimental settings are described in Appendix C.

Baselines. To ensure fair evaluations, we selected baselines that do not necessitate an annotated dataset for training. We examine three word-based jargon detection methods (**Word2Vec** [23], **CantReader** [48], and **Zhu et al.** [54]) as baselines to evaluate their effectiveness in moderating real user-generated conversations. Unlike JEDIS, these methods generate a list of drug jargon after performing analysis on the *corpus-level*. Therefore, the models are trained on the entire corpus. For evaluation, a target word is marked as drug jargon if it appears in the jargon list. Additionally, we present five context-based baselines (**PETD** [15], **MLM *(w/o pretrain)***, **MLM**, **DarkBERT** [11], and **GPT4o-mini** [28]) that inspect *sentence-level* context to determine if a word is used as drug jargon. We also evaluate **variations of JEDIS** to investigate the importance of each component in drug jargon detection. Detailed explanations of baselines are described in Appendix B

5.2 Drug Jargon Detection Results

Word-based vs Context-based detection. The results of our jargon detection experiments are summarized in Table 3. In both datasets, JEDIS achieved the highest F1-scores. The word-based baselines generally have high precision but suffer from low recall. The context-based approaches scored higher F1-scores overall, suggesting they achieved a better tradeoff in finding drug jargon.

Word-based approaches performed better on the Silk Road Forum dataset than on the Reddit Drug dataset, suggesting that the two datasets might be different in the distribution of more difficult jargon terms (such as euphemisms). Word-based approaches are expected to have high precision since many recognized drug terms are used only to refer to drugs. However, their low recall scores strongly suggest that many jargon terms can still evade detection from such systems. The two euphemism-aware baselines, CantReader and Zhu et al.'s model, achieved differing results. Compared to other word-based models, CantReader tends to have low recall, while Zhu et al.'s approach sacrifices precision for better recall. The experiments suggest Zhu et al.'s approach is able to detect more drug contexts, but has more false positives. The word-based baselines achieved fairly high precision scores, but euphemisms prevent these methods from reaching perfect precision. The effect of euphemisms in detection is further discussed in Section 5.3, where we further investigate the words selected by the word-based baselines.

The context-based methods that explicitly consider contexts generally performed better at recall and F1-scores. The one exception is PETD, which was specifically designed to detect euphemisms only. Its limited performance, especially in recall, highlights the necessity for more precise and comprehensive detection method for effective drug content moderation. Further pretraining the MLM model on the corpora (MLM *(w/o pretrain)* vs MLM) improved performance by a considerable amount. This suggests that language models require further pretraining on the domain of drug discussions in order to adapt to the lexical and stylistic characteristics of the domain. This finding is consistent with trends of state-of-the-art models struggling with user-generated texts. However, in the case of DarkBERT, despite not being pretrained on the target corpora, its knowledge of illegal terms obtained from extensive dark web data enabled its detection performance to be comparable to the MLM baseline, particularly in the Reddit Drug dataset. A notable observation is that GPT4o-mini with detailed prompts yields high recall and detection coverage. However, its indiscriminate detection with substantially low precision indicates that even with specifically designed prompts and guidelines (Figure 5), there is still a need for a method carefully tailored for drug jargon detection. Although some context-based baselines achieve high performance, their results are not as precise and generalizable as our delexicalized approach. JEDIS improved performance with its specialized architecture that ensembles the knowledge from both contexts and words.

In terms of detection coverage, JEDIS exhibited a significant capability in detecting a variety of drug jargon terms, even with satisfactory precision. Specifically, among 178 and 169 unique jargon terms from each dataset, JEDIS successfully detected an average of 4.59 times (101 vs 22) more unique terms in the Reddit Drug dataset and 3.86 times (113 vs 29.3) more in the Silk Road Forum dataset compared to the word-based baselines. This underscores JEDIS's superior detection coverage, which significantly exceeds that of existing word-based moderation.

Table 4: Error analysis of each word-based approach. False positives are colored red and false negatives are colored blue.

	Example sentences	Limitations	Example words
Word2Vec [23]	I have also heard that smoking crack with weed is an amazing experience .	Cannot find euphemisms	crack, speed, weed
CantReader [48]	I've just tried it before and its niceeeeee .	Sensitive to typos	comfused, niceeeeee
Zhu et al. [54]	Same thing happened with 10mg valium + 40mg meth .	limited to BERT vocabulary	add, met, valium

JEDIS ablation study. We examine the impact of JEDIS components by comparing the performance of its variations. The results of JEDIS (*w/o* Neg_{STMS}) suggest that the utilization of noisy negative data from STMS resulted in a significant improvement in precision. At the same time, we observed that JEDIS (*w/o* Neg_{STMS}) achieved fairly high recall on both datasets. This can be attributed to the unknown drug jargon in STMS, being labeled as negatives due to being outside of the seed terms during the distant supervision setting. The effect of pretraining was significant, as JEDIS shows performance improvements in both precision and recall by pretraining. The enhancement in recall was especially notable, suggesting that domain knowledge obtained from pretraining on drug discussions aided in detecting more challenging and diverse drug jargon. When comparing JEDIS and JEDIS (*w/o word*), we find that utilizing word information led to improvements in precision. However, when comparing two non-pretrained models (JEDIS (*w/o pretrain*) and JEDIS (*w/o pretrain&word*)), using word information leads to a decrease in recall and only a slight improvement in precision. This suggests that when utilizing word embeddings from an off-the-shelf BERT model, which lacks adequate knowledge of drug contexts, the embeddings can actually hinder the detection of a diverse range of drug jargon.

From the results, we suggest the effectiveness of context-based approaches in content moderation. We show JEDIS has superior performance on the jargon detection task compared to other baselines. We also demonstrate the effectiveness of JEDIS in a cross-domain setting in Appendix F.

5.3 Qualitative Analysis

We further conduct qualitative analysis to investigate the limitations of each moderation approach. In this task, we compare the generated lists of detected drug jargon by each approach. We compare JEDIS and the four word-based methods trained in Section 5.1.

Drug jargon list construction. We construct a list of words by applying JEDIS to all candidates (alphanumeric nouns) in the *entire corpus*. To score a word candidate, we count both prediction frequency F_{pred} (# positive predictions) and prediction ratio R_{pred} (# positive predictions divided by # total predictions made) of the word. Our score is given by the following formula: $Score = F_{pred} * {R_{pred}}^n$. The weighing parameter n adjusts sensitivity to frequently occurring words. We set $n = 2$ and extracted the top 100 words. For word-based methods, we also select the top 100 words.

Analysis. For each word, we manually investigate if the word was used as drug jargon in the corpus (using the guidelines established before). We present the top 100 extracted jargon list of each approach in our repository. We also identify cases of failure of the baseline approaches and present examples in Table 4.

The Word2Vec methods showed high performance in precision. However, Word2Vec may not be able to effectively identify euphemisms since it assumes a single meaning for each word. We find that the Word2Vec list missed many euphemisms found by other approaches. For instance, Word2Vec could not identify *weed*, *speed*, *lucy*, *spice*, and *crack*, which were all identified by both JEDIS and Zhu et al.'s model. Therefore, adversaries can easily exploit Word2Vec's weakness by using common words as euphemisms.

CantReader showed low overall precision with many false positives. CantReader is designed to find euphemisms by finding words that are used differently across corpora. However, with the exception of *weed*, it was also not able to find the above euphemisms that Word2Vec failed to find. The CantReader word list reveals that many false positives came from misspelled non-jargon words, such as *superawesome*, *comfused*, and *niceeeeee*. By design, the system gives infrequent words high scores since they may be characteristic to the target corpus. However, this suggests it is unsuitable to use with user-generated texts, which often contain typos.

Zhu et al.'s approach is based on making a pretrained BERT model that predicts how likely each word is able to fit a context. Although this model is capable of finding many euphemistic terms, words outside of BERT's limited wordpiece vocabulary cannot be detected by this approach. Terms like *valium*, *ambien*, *diazepam*, *oxy*, and *focalin* were terms found by other methods, but were not found by Zhu et al.'s model because they are not in BERT's vocabulary. This limitation also causes false positives. BERT splits out-of-vocabulary words into wordpieces, so unknown words (like *meth*) will be represented as a combination of two words (*met* and *##h*). The false positives *met*, *add*, and *mx* may indicate *meth*, *adderall*, and *mxe*. This highlights the limitation that the model cannot predict out-of-vocabulary words correctly, even when the model has sufficiently learned on them.

JEDIS judges words mainly by how their contexts appear to be. Therefore, a common type of error was from legal drugs that are often discussed in similar ways to illicit drugs, such as *alcohol* and *caffeine*. However, these errors were common in other approaches as well, suggesting it is a general linguistic limitation, which we elaborate on in Section 6.1. JEDIS, unlike other baselines, had no significant systematic failures in recall. Qualitative analysis suggests JEDIS exhibits resilience to a variety of evasion strategies, while maintaining extensive detection coverage.

Table 5: Jargon detection results on the Reddit Drug dataset. JEDIS$_{unamb}$ is trained on unambiguous data, while JEDIS$_{control}$ uses randomly sampled control data.

	# Positive data	Precision	Recall	F1-score
JEDIS$_{original}$	329,573	0.6659	0.6197	0.6419
JEDIS$_{unamb}$	46,325	0.7949	0.2774	0.4113
JEDIS$_{control}$	46,325	0.5966	0.6219	0.6090

6 Discussion

6.1 Linguistic Challenges

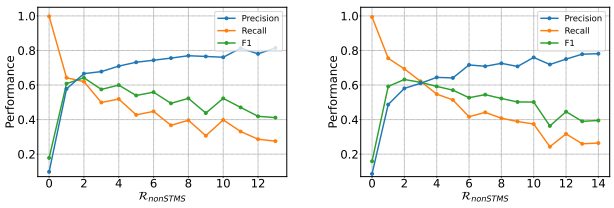
We have argued for a delexicalized approach to automate content moderation. However, an entirely delexicalized detection has its limitations. Previous work noted that not all contexts of drug jargon are indicative of drug activity [54]. For instance, the sentence "Oh boy do I love ketamine" from our corpus features the drug *ketamine* in a non-indicative context. A challenge for a detection system is to understand that the context "Oh boy do I love [MASK]" can potentially accommodate both drug jargon (*ketamine*) and non-drug words (*Fridays*). For these ambiguous cases, lexical information of the masked word is required for an accurate prediction. JEDIS successfully addresses these cases by ensembling delexicalized prediction with lexically informed prediction.

Another challenge is to make the distinction between proper and improper usages of legal drugs. Legal substances such as *caffeine* or *alcohol* may be consumed irresponsibly ("I'm already addicted to *alcohol*"), and can cause false positives. While these false positives can be filtered relatively easily with an "allowlist", the problem extends to prescription drugs. Prescription drugs such as *vyvanse* have both legal and abusive usages. Some contexts may explicitly imply legal usage ("I've been prescribed Vyvanse for the past 2 years"), but many discussions on these drugs ("Will *Vyvanse* sabotage the [workout] progress I'm making") may not contain enough information to evaluate the legality of usage. It could be a further challenge to identify sentences of ambiguous legality. Whether to permit or moderate these cases could be a matter of policy.

6.2 Effects of JEDIS Configurations

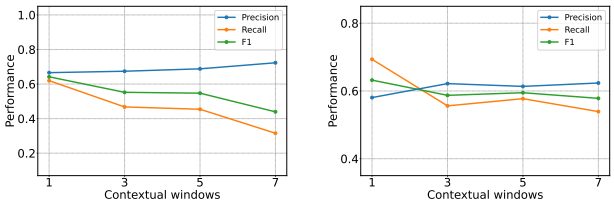
Effective moderation using JEDIS requires careful configurations. Thus, we analyze various configuration aspects and find insights into how they affect the overall performance. We also discuss the potential application of JEDIS to other fields in Appendix G.

Effect of ambiguous contexts. As discussed in Section 6.1, drug jargon can emerge in ambiguous contexts, which might complicate their usage as training data by introducing confusing samples. Addressing this, Zhu et al. [54] suggested a filtering technique for such contexts and demonstrated enhanced performance when leveraging only unambiguous contexts for training. To examine this effect on JEDIS, we adopt Zhu et al.'s technique, trimming our potentially ambiguous training data to only unambiguous samples. As a control, we also create a dataset of equivalent size from the unfiltered data. With the same hyperparameters for negative training data, we assess jargon detection performance on these new training datasets, as detailed in Table 5. The model trained on only unambiguous data suffered a huge drop in overall performance, while the control



(a) Reddit Drug (b) Silk Road Forum

Figure 3: Jargon detection performance with varying values of $\mathcal{R}_{nonSTMS}$. $\mathcal{R}_{STMS}$ is set to 5.



(a) Reddit Drug (b) Silk Road Forum

Figure 4: Jargon detection performance with varying contextual window sizes.

model only marginally dropped. When trained without ambiguous data, the model was able to achieve improvements in precision, but drastically lost recall. The results suggest that the model trained only on clear-cut positive data cannot generalize fully to the diverse contexts that drug jargon appears. We hypothesize that a model that understands the potential illicitness of ambiguous contexts is generally better than one that disregards them completely.

Effect of negative ratios. If we rely too heavily on distant supervision (i.e., label every other context as negatives), jargon outside the seed term list will be falsely marked as negatives. Therefore, we select only a portion of the available context to use as negatives by choosing a hyperparameter $\mathcal{R}_{nonSTMS}$. Figure 3 illustrates JEDIS's performance with varying $\mathcal{R}_{nonSTMS}$ values, which can control the influence of potentially noisy negative labels. Models trained without $Neg_{nonSTMS}$ exhibit low precision and high recall, indicating a bias towards positive predictions. While negative training sentences are essential, an excess can degrade performance, likely from noisy label introduction. A rise in $\mathcal{R}_{nonSTMS}$ boosts precision but diminishes recall, indicating a tradeoff determined by the volume of negative training data. For both datasets, having $\mathcal{R}_{nonSTMS} = 2$ produced the best F1-score. However, $\mathcal{R}_{nonSTMS}$ can further be tuned depending on whether a model is designed to be *precision-sensitive* or *recall-sensitive*.

Effect of contextual windows. JEDIS conducts detection at the single-sentence level. To explore the impact of broader contexts, we expanded the contextual window to include up to three additional sentences on either side of the target sentence, increasing the window sizes to encompass 3, 5, and 7 sentences in total. As shown in Figure 4, this expansion improved precision by providing more contextual clues for more accurate detection. However, we observed a significant reduction in recall and F1-score, primarily

Table 6: Jargon detection results on Reddit Drug evaluation dataset with varying numbers of seed terms.

$\text{JEDIS}_{\#\text{ seed terms}}$	Precision	Recall	F1-score
JEDIS_{23}	0.6039	0.3445	0.4387
JEDIS_{46}	0.6659	0.6197	0.6419
JEDIS_{70}	0.6781	0.6242	0.6503

due to the amplification of false negatives. This issue arises from the use of noisy negatives through distant supervision, which becomes more pronounced with larger windows. Moreover, larger windows increased computational overhead during training and inference. Given these factors, we concluded a single-sentence window is the optimal approach in configuring JEDIS.

Effect of seed terms and iterative use of JEDIS. We examine the effect of drug seed terms on JEDIS by repeating experiments with reduced and extended seed term lists. First, we train a model with a reduced seed term list of 23 words (randomly selected from the original list of 46 terms). Then, we train another model with an extended 64-word seed term list, which was created by adding 18 new words that JEDIS correctly found. The results are shown in Table 6. JEDIS trained on the reduced list achieves a worse result than the original, especially in recall. This suggests JEDIS is indeed sensitive to the seed terms used to create distant supervision. However, the model trained on the reduced list still manages to reach a reasonable performance. It could be used to bootstrap a larger list of seed terms by discovering new jargon. Although human effort is required to identify quality terms, this can facilitate the development of a seed list comparable to the original, achieving similar performance. Similarly, the original model can be used to iteratively increment the seed term list with newly detected words. When JEDIS was trained on the extended list, we observe that the performance increased by a relatively small amount. This suggests JEDIS can be improved with a larger list of seed terms, but with diminishing returns. This iterative updating process illustrates JEDIS's capacity to adapt to evolving landscapes, continuously refining its performance by integrating new seed terms.

7 Conclusion

Automatic content moderation of drugs on social media platforms is an important but challenging task. We present the drug jargon detection framework JEDIS, which can handle unseen terms and euphemisms more effectively than previously proposed methods. JEDIS achieves this through a novel delexicalized distant supervision training method on an unannotated corpus. Our evaluations on two manually annotated datasets show that JEDIS outperforms state-of-the-art baselines. Through qualitative analysis, we show that JEDIS also has superior coverage and robustness.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS- 2023-00215700, Trustworthy Metaverse: blockchain-enabled convergence research, 50%), and S2W Inc. through financial support and technical assistance (50%).

References

[1] Drug Enforcement Administration. 2018. Slang Terms and Code Words: A Reference for Law Enforcement Personnel. https://www.dea.gov/sites/default/files/2018-07/DIR-022-18.pdf. Accessed: 2022-10-26. (2018).

[2] Gwern Branwen et al. 2015. Dark Net Market archives, 2011-2015. https://www.gwern.net/DNM-archives. dataset. Accessed: 2022-10-26. (July 2015). https://www.gwern.net/DNM-archives.

[3] Edward J. Cone. 2006. Ephemeral profiles of prescription drug and formulation tampering: evolving pseudoscience on the internet. *Drug and Alcohol Dependence*, 83, S31–S39. Drug Formulation and Abuse Liability. DOI: https://doi.org/10.1016/j.drugalcdep.2005.11.027.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[5] Centers for Disease Control and Prevention. 2022. Drug Overdose Deaths in the United States, 2001–2021. https://www.cdc.gov/nchs/products/databriefs/db457.htm. Accessed: 2023-02-07. (2022).

[6] Lorenzo Franceschi-Bicchierai. 2018. Reddit Bans Subreddits Dedicated to Dark Web Drug Markets and Selling Guns. https://www.vice.com/en/article/ne9v5k/reddit-bans-subreddits-dark-web-drug-markets-and-guns. Accessed: 2022-10-26. (2018).

[7] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, (July 2020), 8342–8360. DOI: 10.18653/v1/2020.acl-main.740.

[8] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, cucks, and god emperor trump: a measurement study of 4chan's politically incorrect forum and its effects on the web. In *ICWSM*.

[9] Felipe Hoffa. 2015. 1.7 billion reddit comments loaded on BigQuery. https://www.reddit.com/r/bigquery/comments/3cej2b/17_billion_reddit_comments_loaded_on_bigquery/. Accessed: 2022-10-25. (2015).

[10] Chuanbo Hu, Minglei Yin, Bin Liu, Xin Li, and Yanfang Ye. 2021. Detection of illicit drug trafficking events on instagram: a deep multimodal multilabel learning approach. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3838–3846.

[11] Youngjin Jin, Eugene Jang, Jian Cui, Jin-Woo Chung, Yongjae Lee, and Seungwon Shin. 2023. Darkbert: a language model for the dark side of the internet. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7515–7533.

[12] Youngjin Jin, Eugene Jang, Yongjae Lee, Seungwon Shin, and Jin-Woo Chung. 2022. Shedding new light on the language of the dark web. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, (July 2022), 5621–5637. DOI: 10.18653/v1/2022.naacl-main.412.

[13] Liang Ke, Xinyu Chen, and Haizhou Wang. 2022. An unsupervised detection framework for chinese jargons in the darknet. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (WSDM '22). Association for Computing Machinery, Virtual Event, AZ, USA, 458–466. ISBN: 9781450391320. DOI: 10.1145/3488560.3498469.

[14] Hannah Kirk, Bertie Vidgen, Paul Rottger, Tristan Thrush, and Scott Hale. 2022. Hatemoji: a test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, (July 2022), 1352–1368. DOI: 10.18653/v1/2022.naacl-main.97.

[15] Patrick Lee, Martha Gavidia, Anna Feldman, and Jing Peng. 2022. Searching for PETs: using distributional and sentiment-based methods to find potentially euphemistic terms. In *Proceedings of the Second Workshop on Understanding Implicit and Underspecified Language*. Valentina Pyatkin, Daniel Fried, and Talita Anthonio, (Eds.) Association for Computational Linguistics, Seattle, USA, (July 2022), 22–32. DOI: 10.18653/v1/2022.unimplicit-1.4.

[16] Jiawei Li, Qing Xu, Neal Shah, Tim K Mackey, et al. 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, 6, e13803.

[17] Tim K Mackey and Janani Kalyanam. 2017. Detection of illicit online sales of fentanyls via twitter. *F1000Research*, 6.

[18] Tim K Mackey, Janani Kalyanam, Takeo Katsuki, and Gert Lanckriet. 2017. Twitter-based detection of illegal online sale of prescription opioid. *American journal of public health*, 107, 12, 1910–1915.

[19] Rijul Magu, Kshitij Joshi, and Jiebo Luo. 2017. Detecting the hate code on social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 11, 1, (May 2017), 608–611. DOI: 10.1609/icwsm.v11i1.14921.

[20] Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, Brussels, Belgium, (Oct. 2018), 93–100. DOI: 10.18653/v1/W18-5112.

[21] Andrei Manolache, Florin Brad, Antonio Barbalau, Radu Tudor Ionescu, and Marius Popescu. 2022. Veridark: a large-scale benchmark for authorship verification on the dark web. *Advances in Neural Information Processing Systems*, 35, 15574–15588.

[22] Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. *arXiv preprint arXiv:2109.05003*.

[23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[24] Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, (June 2013), 777–782. https://aclanthology.org/N13-1095.

[25] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, Suntec, Singapore, (Aug. 2009), 1003–1011. https://aclanthology.org/P09-1113.

[26] United Nations. [n. d.] Break the link between illicit drugs and social media: un-backed report | un news. (). https://news.un.org/en/story/2022/03/1113642.

[27] National Institutes of Health. 2022. Overdose Death Rates. https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates. Accessed: 2022-11-08. (2022).

[28] OpenAI. 2024. Gpt4omini. https://platform.openai.com/docs/models/gpt-4o-mini. Accessed: 2024-12-04. (2024).

[29] Jose Pagliery. 2013. FBI shuts down online drug market Silk Road. https://money.cnn.com/2013/10/02/technology/silk-road-shut-down/index.html. Accessed: 2022-10-26. (2013).

[30] Yiyue Qian, Yiming Zhang, Yanfang (Fa Ye, and Chuxu Zhang. 2021. Distilling meta knowledge on heterogeneous graph for illicit drug trafficker detection on social media. In *Advances in Neural Information Processing Systems*. M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, (Eds.) Vol. 34. Curran Associates, Inc., 26911–26923. https://proceedings.neurips.cc/paper/2021/file/e234e195f3789f05483378c397db1cb5-Paper.pdf.

[31] Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, (July 2018), 2137–2147. DOI: 10.18653/v1/P18-1199.

[32] Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, (Apr. 2017), 1171–1182. https://aclanthology.org/E17-1110.

[33] Avik Ray, Yilin Shen, and Hongxia Jin. 2019. Iterative delexicalization for improved spoken language understanding. *arXiv preprint arXiv:1910.07060*.

[34] Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. English. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. European Language Resources Association (ELRA), Marseille, France, (May 2020), 137–143. ISBN: 979-10-95546-56-6. https://aclanthology.org/2020.trac-1.22.

[35] Sanna Rönkä and Anu Katainen. 2017. Non-medical use of prescription drugs among illicit drug users: a case study on an online drug forum. *International Journal of Drug Policy*, 39, 62–68. DOI: https://doi.org/10.1016/j.drugpo.2016.08.013.

[36] Vageesh Saxena, Nils Rethmeier, Gijs van Dijck, and Gerasimos Spanakis. 2023. VendorLink: an NLP approach for identifying & linking vendor migrants & potential aliases on Darknet markets. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, (Eds.) Association for Computational Linguistics, Toronto, Canada, (July 2023), 8619–8639. DOI: 10.18653/v1/2023.acl-long.481.

[37] Dominic Seyler, Wei Liu, XiaoFeng Wang, and ChengXiang Zhai. 2021. Towards dark jargon interpretation in underground forums. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II*. Springer-Verlag, Berlin, Heidelberg, 393–400. ISBN: 978-3-030-72239-5. DOI: 10.1007/978-3-030-72240-1_40.

[38] Dominic Seyler, Wei Liu, Yunan Zhang, XiaoFeng Wang, and ChengXiang Zhai. 2021. Darkjargon.net: a platform for understanding underground conversation with latent meaning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '21). Association for Computing Machinery, Virtual Event, Canada, 2526–2530. ISBN: 9781450380379. DOI: 10.1145/3404835.3462801.

[39] Sylvester Shan and Ramesh Sankaranarayana. 2020. Behavioral profiling of darknet marketplace vendors. (2020).

[40] Donghyun Son, Byounggyu Lew, Kwanghee Choi, Yongsu Baek, Seungwoo Choi, Beomjun Shin, Sungjoo Ha, and Buru Chang. 2023. Reliable decision from multiple subtasks through threshold optimization: content moderation in the wild. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (WSDM '23). Association for Computing Machinery, Singapore, Singapore, 285–293. ISBN: 9781450394079. DOI: 10.1145/3539597.3570439.

[41] Kyle Soska and Nicolas Christin. 2015. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *Proceedings of the 24th USENIX Conference on Security Symposium* (SEC'15). USENIX Association, Washington, D.C., 33–48. ISBN: 9781931971232.

[42] Christophe Soussan and Anette Kjellgren. 2014. Harm reduction and knowledge exchange—a qualitative analysis of drug-related internet discussion forums. *Harm reduction journal*, 11, 1, 1–9.

[43] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21) Article 341. Association for Computing Machinery, Yokohama, Japan, 14 pages. ISBN: 9781450380966. DOI: 10.1145/3411764.3445092.

[44] Sandeep Suntwal, Mithun Paul, Rebecca Sharp, and Mihai Surdeanu. 2019. On the importance of delexicalization for fact verification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, (Nov. 2019), 3413–3418. DOI: 10.18653/v1/D19-1340.

[45] Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, (June 2021), 576–587. DOI: 10.18653/v1/2021.naacl-main.48.

[46] Hao Yang, Xiulin Ma, Kun Du, Zhou Li, Haixin Duan, Xiaodong Su, Guang Liu, Zhifeng Geng, and Jianping Wu. 2017. How to learn klingon without a dictionary: detection and measurement of black keywords used by the underground economy. In *2017 IEEE Symposium on Security and Privacy (SP)*, 751–769. DOI: 10.1109/SP.2017.11.

[47] Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised NER with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, (Aug. 2018), 2159–2169. https://aclanthology.org/C18-1183.

[48] Kan Yuan, Haoran Lu, Xiaojing Liao, and XiaoFeng Wang. 2018. Reading thieves' cant: automatically identifying and understanding dark jargons from cybercrime marketplaces. In *27th USENIX Security Symposium (USENIX Security 18)*. USENIX Association, Baltimore, MD, (Aug. 2018), 1027–1041. ISBN: 978-1-939133-04-5. https://www.usenix.org/conference/usenixsecurity18/presentation/yuan-kan.

[49] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is gab: a bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, 1007–1014.

[50] Yiming Zhang, Yiyue Qian, Yujie Fan, Yanfang Ye, Xin Li, Qi Xiong, and Fudong Shao. 2020. Dstyle-gan: generative adversarial network based on writing and photography styles for drug identification in darknet markets. In *Annual Computer Security Applications Conference*, 669–680.

[51] Yiming Zhang et al. 2019. Your style your identity: leveraging writing and photography styles for drug trafficker identification in darknet markets over attributed heterogeneous information network. In *The World Wide Web Conference*, 3448–3454.

[52] Kangzhi Zhao, Yong Zhang, Chunxiao Xing, Weifeng Li, and Hsinchun Chen. 2016. Chinese underground market jargon analysis based on unsupervised learning. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, 97–102. DOI: 10.1109/ISI.2016.7745450.

[53] Wanzheng Zhu and Suma Bhat. 2021. Euphemistic phrase detection by masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, (Nov. 2021), 163–168. DOI: 10.18653/v1/2021.findings-emnlp.16.

[54] Wanzheng Zhu, Hongyu Gong, Rohan Bansal, Zachary Weinberg, Nicolas Christin, Giulia Fanti, and Suma Bhat. 2021. Self-supervised Euphemism Detection and Identification for Content Moderation. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 229–246.

Appendices

A Annotation guidelines.

We provide the annotation guidelines developed after the pilot session. The detailed guidelines with examples are as follows.

- Objective: Find words used as drug jargon in the given sentences.
- When the word "drug(s)" is used to indicate an instance of the drug, annotate it as drug jargon. (e.g., I was high on **drugs** at the time.)
- When the word "drug(s)" is used to describe just the concept of drugs, annotate it as NOT drug jargon. (e.g., I dismissed it as just a party **drugs**.)
- When a word indicating the form of a drug is used as a substitute for drug words, annotate it as drug jargon. (e.g., The **bars** can get you pretty high.)
- When a word indicating the form of a drug is used alongside drug words, annotate it as NOT drug jargon. (e.g., He had a bunch of Xanax **bars**.)
- When a word refers to a legal drug, but the context is discussing it for abusive purposes, annotate it as drug jargon. (e.g., You can get a great stim high from **propylhexedrine** in otc inhalers.)

B Baselines

For all relevant baseline models, hyperparameter K was empirically tuned to achieve *the best performance*.

- **Word2Vec** [23] Word2Vec learns vector representation of words. We employ the CBOW mode, which trains the model by predicting the middle word based on its surrounding context. We use the most_similar function of the Gensim library implementation of Word2Vec to obtain the most similar words to the 46 seed terms according to cosine similarity. This process generated jargon lists of 183 and 175 words from the Reddit Drug corpus and the Silk Road Forum corpus, respectively.
- **CantReader** [48] CantReader leverages the difference in word usage across different corpora (dark, legitimate, and reputable) to find euphemistic jargon. We use our forum corpora as the dark corpus to extract a drug jargon list. The list is constructed by taking the top K words produced by CantReader. We set K as 120 for the Reddit Drug dataset and 220 for the Silk Road Forum dataset.
- **Zhu et al.** [54] A BERT-based MLM model is pretrained on the target corpus. Then, sentences containing drug seed terms are filtered to keep only sentences with informative drug contexts. From the informative sentences, the *drug seed terms* are masked. The MLM model predicts words that can replace the mask tokens. The system produces a jargon list by sorting accumulated MLM probabilities. It is capable of finding both euphemistic and non-euphemistic drug jargon. We take top K words to create the banlist. We set K as 40 and 50 for each dataset, respectively.
- **PETD** [15] PETDetection finds potentially euphemistic terms using a combination of distributional similarities and sentiment-based metrics. It involves extracting phrase candidates from the sentence, filtering them based on their relevance to the topic (i.e., *drug*), paraphrasing to identify sentiment shifts, and ranking the phrases. For each sentence, we mark a word as the drug jargon when it passes all the stages (extraction, filtering, paraphrasing, and ranking).
- **MLM *(w/o pretrain)*** We newly present another baseline that uses MLM probability in the different manner with Zhu et al. Instead of masking known drug words, we mask the *target word* to be inspected. If a drug seed term appears in the top K words predicted by the MLM, the masked word is predicted as a drug jargon in the sentence. We use the BERT-base-uncased as the MLM model, which has been pretrained on generic web texts. We set K as 100 for both datasets.
- **MLM** The same approach as above, but we further pretrain the MLM model on the target corpus. We set K as 50 for the Reddit Drug dataset and 30 for the Silk Road Forum dataset.
- **DarkBERT** [11] DarkBERT is a RoBERTa-based language model pretrained on Dark Web data. Leveraging the domain-specific knowledge acquired from Dark Web data, it demonstrates superior performance in detecting threat keywords, including illicit drug terms. Following the methodology outlined in the original paper [11], if a drug seed term appears in the top K related words predicted by the MLM, the masked word is identified as drug jargon in the sentence. We set K to 40 for both datasets.
- **GPT4o-mini** [28] GPT4o-mini can be utilized for content moderation through prompt learning. As an assistant to content moderators, we use specifically designed prompts (as shown in Figure 5) and feed them with sentences where the inspected words are highlighted with [SP] tokens. This baseline demonstrates the utility of recent large language models and prompt learning in the context of drug content moderation.
- **JEDIS ablation variations** Settings of JEDIS without using negative data from STMS (*w/o* Neg_{STMS}), pretraining BERT on the target domain (*w/o pretrain*), word attribute module (*w/o word*), or both pretraining and word attribute module (*w/o pretrain&word*). These baselines suggest the importance of information obtained from each component for drug jargon detection.

C Experimental Settings

Evaluations were done on a machine with two Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz and two NVIDIA GeForce RTX 4090s. For the BERT models used in *Classification* modules, we used the bert-base-uncased model with the embedding dimension of 768, max length of 128, batch size of 32, learning rate of 1e-5, and warmup ratio of 0.1. For further pretraining, we ran run_mlm.py on HuggingFace with max sequence length of 512, validation split percentage of 10, batch size of 32 train epochs of 3. Then, we trained the classification models with the AdamW optimizer and cross entropy loss for a maximum of 10 epochs with early stopping enabled. We used Python version 3.10 for all implementations.

We also train the baseline models as follows: For Word2Vec baselines, min count of 10, vector size of 100, window size of 10,

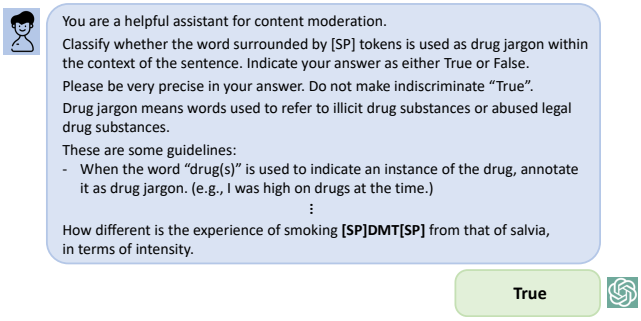


Figure 5: Prompts fed into GPT4o-mini. The definition of drug jargon in Section 2.1 and the annotation guidelines in Appendix A are used.

epochs of 10 are used as parameters. For Zhu et al.'s approach, we follow the parameters described in the original paper. For the MLM baselines, we ran run_mlm.py on HuggingFace using bert-base-uncased with max sequence length of 512, validation split percentage of 10 batch size of 32 train epochs of 3. For PETD, we use the topic word as *drug* [7] and follow the other parameters described in the original paper.

D Ethical Considerations

Because our work is closely related to user-generated content, we consider the ethical concerns during our study. First, we use only publicly released data accessible to anyone, as mentioned in Section 4.1. Other existing works use the same data for clustering drug vendors in darknet markets [21, 36, 39, 50, 51] or finding drug jargon in social media [53, 54]. Second, we anonymized or removed the information that could possibly lead to any harm, such as usernames, email addresses, and PGP keys. Finally, we will open our annotated datasets to only researchers who fill out a request form with clear research purposes and verifiable identities. We believe that the contributions of our work on automatic content moderation and the application of NLP techniques could outweigh the potential risks derived from our study.

E How to address potential subjectivity and bias in annotation

We constructed evaluation datasets through manual annotation. However, this process may introduce potential subjectivity or bias from the annotators, which could affect the quality of the constructed dataset. To mitigate these effects, we implemented the following process: First, we recruited two expert annotators from a cybersecurity company specializing in monitoring malicious online activities. They possess substantial knowledge of drug jargon and high expertise in related tasks. Second, as described in Section 4.2, we conducted a pilot session using the first 100 sentences of each dataset. Subsequently, we developed detailed annotation guidelines with examples, as outlined in Appendix A. As shown in Table 2, this resulted in a very high inter-annotator agreement. Finally, in cases of disagreement between annotators, thorough discussions were held to reach a consensus. Through these steps, we made careful efforts to minimize potential bias and subjectivity.

[7] We also experimented with using our 46 drug seed terms as topic words, but this resulted in poorer performance.

Table 7: Evaluation results on cross domain performance. $T_{A \to B}$ denotes performance of JEDIS trained on corpus A and evaluated on corpus B.

Domain Transfer	Precision	Recall	F1-score
$T_{Reddit \to Reddit}$	0.6659	0.6197	0.6419
$T_{SR \to Reddit}$	0.5868	0.5973	0.5920
$T_{SR \to SR}$	0.5805	0.6934	0.6320
$T_{Reddit \to SR}$	0.5616	0.5497	0.5556

F Cross Domain Experiments

To investigate whether JEDIS can perform well in unseen domains, we conduct jargon detection using the model trained on the other dataset. The experimental results are shown in Table 7. In both datasets, JEDIS trained on the target dataset produced higher performance. However, the cross-domain models also produced reasonably high performance comparable to the pretrained MLM approach. This suggests that the languages of the two datasets do differ, but knowledge learned from one domain could be transferred to the target domain. The results suggest that while training the classification model on the target domain remains ideal, our approach has the ability to generalize to unseen domains that do not have sufficient training data (such as emerging social media sites).

G Application to Other Fields

Content moderation can encompass other types of forbidden content, such as discussion on weapons, sexually explicit, or hate speech. While JEDIS was designed for drug jargon detection, its foundational principles permit broader applications. However, the success of JEDIS rests on the distinctiveness of jargon contexts from non-jargon. Illicit drug terms have specific linguistic features and contextual cues that might not be as evident in other domains. For JEDIS to be effectively repurposed for other areas, the target jargon should have contexts as distinguishable as drug terms. Thus, while the potential exists for JEDIS's adaptation to domains with distinct linguistic characteristics, its efficacy in such extensions warrants rigorous future exploration.

Pilot experiment on sexually explicit domain. To evaluate the potential for adapting JEDIS to other domains, we conducted a pilot study focusing on the moderation of sexually explicit discussions. The training corpus comprised 1,000,000 sentences extracted from user discussions on the GAB platform [49], with 43 sexual seed words, which were manually identified, serving as seed terms. Given the absence of the annotated dataset for evaluation, an additional set of 10 sexual words was used as test terms to assess JEDIS's performance in detecting these terms within the test corpus. The outcomes demonstrated a precision of *0.6585*, a recall of *0.5562*, and an F1-score of *0.6030* [8]. These results indicate that JEDIS holds promise for application in domains beyond its original design, even when not optimized through hyperparameter adjustments.

[8] The experimental results do not reflect precise accuracy due to the test set's construction, which relied solely on the presence of test terms without annotation.