Few-Label SetFit for C2C Online Marketplace Listings: Multi-Label Classification and Entity Extraction for Identifying Potentially Illicit Ads

Pablo Rivas[1] **Jorge Yero Salazar**[1] **Bikram Khanal**[1] **Javier S. Turek**[2]
[1] Department of Computer Science, Baylor University, Waco, TX, USA
[2] EarthDynamics.AI, Beaverton, OR, USA
{pablo_rivas, jorge_yero1, bikram_khanal1}@baylor.edu
javier@earthdynamics.ai

Abstract

The proliferation of illicit online advertisements poses a growing challenge for law enforcement and public safety, demanding methods that are both accurate and data-efficient. This paper investigates the use of `SetFit`, a prompt-free few-shot framework for fine-tuning sentence transformers, for detecting risk signals in online advertisements. We study two domains of high societal impact: commercial sex advertisements (CSAs) associated with suspected human trafficking activity (HT) and suspected stolen-parts marketplace listings (SCP). For CSAs, we cast the problem as a multi-label post-classification task and evaluate different sentence-aggregation strategies; our best approach, averaging across all sentences, achieves an F1 of **0.783**, surpassing a GPT-2 baseline trained with more data and significantly larger parameter counts. In contrast, when applied to token-level named-entity recognition for suspected stolen-parts marketplace listings (SCP), SetFit underperforms (F1 = 0.242) relative to GPT-2 (F1 = 0.337), exposing a structural limitation in adapting sentence-level embeddings to fine-grained extraction. These findings demonstrate that SetFit is highly effective for low-resource post classification in sensitive domains, while also motivating new research directions for extending few-shot methods to entity-level modeling.

1 Introduction

Illegal online ads hide within large volumes of normal content and span many forms of harm: sexual exploitation, illicit trade, scams, and abuse. Recent work shows fast progress in detecting harmful content across text and other media, with systems that run in real time and cover many languages [Koreddi et al., 2025, Bensoltane and Zaki, 2024]. Broader reviews also show a shift toward end-to-end learning for safety tasks, while calling out open issues in robustness, privacy, and evaluation [Yuan et al., 2023]. Beyond general safety, there are focused efforts on crime-linked markets, such as cultural-heritage trafficking and wildlife trade, which require domain context and careful evidence trails [Yuan et al., 2023, Mou et al., 2024]. These trends motivate methods that work with scarce labels, respect privacy, and give useful signals for screening and triage.

We study two high-impact settings where labels are scarce and noise is common: (i) commercial sex advertisements (CSAs) associated with suspected HT activity, and (ii) ads possibly linked to stolen car parts. For human trafficking, we target CSAs that may indicate HT risk, often linked to organized groups [Kristof, 2012]. For stolen car parts, we target suspected

Proceedings of the LatinX in AI Workshop @ NeurIPS-25.

stolen-parts listings within consumer-to-consumer marketplaces [Aniello and Caneppele, 2018]. We do not disclose platform names to comply with the Institutional Review Board (IRB)-approved, ethics-aligned data-collection protocol.

Our pipeline uses two NLP tasks that match field practice. First, multi-label post classification flags risky content and broad categories within a post. Second, named-entity recognition (NER) extracts details (e.g., names, phones, places) that aid triage and case linking. Since annotation is costly and data access is limited, we use a few-shot setup built on sentence embeddings. We adopt SetFit, a prompt-free fine-tuning method for sentence transformers that learns from a small set of labeled pairs and then trains a simple classifier on top [Tunstall et al., 2022]. Sentence Transformers give strong sentence-level representations that work well with contrastive training [Reimers and Gurevych, 2019]. This design also fits privacy rules because it needs fewer labeled examples and can run with light supervision.

From a technical view, our study probes the boundary between sentence-level and token-level learning under low-label regimes. Post classification maps well to sentence embeddings and pairwise training; token-level NER is harder due to short spans, markup artifacts, and code-mixed text. This gap also appears in recent safety systems that do well at post-level screening but still face noise and domain shift at fine-grained extraction [Koreddi et al., 2025, Bensoltane and Zaki, 2024]. We make this tension explicit and test simple, reproducible choices that matter in practice (e.g., sentence aggregation rules for multi-sentence posts), while keeping a clear link to field needs in public safety and policy [Yuan et al., 2023, Mou et al., 2024].

Contributions. (1) A low-label pipeline for crime-linked online ads that pairs multi-label post classification with NER, using SetFit on top of sentence transformers [Tunstall et al., 2022, Reimers and Gurevych, 2019]. (2) A study of sentence aggregation strategies for multi-sentence posts and their impact on few-shot classification. (3) A negative result for token-level NER under the same few-shot regime, with a clear error analysis and paths forward (e.g., span heads, weak labels, and noise-robust loss). (4) An ethics and deployment note: privacy-aware annotation, risk of false alerts, and the need for human oversight when screening sensitive content [Yuan et al., 2023].

An overview of the end-to-end pipeline is shown in Figure 1 and detailed in Section 3.3.

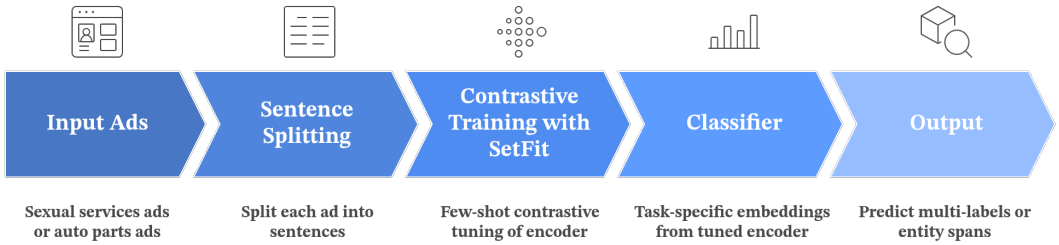

Figure 1: End-to-end pipeline: raw ads are split into sentences, the encoder is fine-tuned with SetFit using contrastive pairs, and the tuned embeddings feed a lightweight head for either multi-label post classification or entity extraction. Mean pooling over sentences yields the best post-level representation in our setup.

2 Background

This section covers three building blocks: multi-label classification, NER, and few-shot sentence-level fine-tuning with SetFit and related methods. We give formal definitions and then place our choices in the context of recent work.

Problem setup. Let a post A consist of sentences $S = \{s_1, \ldots, s_n\}$ with sentence embeddings $h_i \in \mathbb{R}^d$ produced by a SetFit-tuned encoder $E(\cdot)$. For CSA classification (HT risk), we model multi-label prediction over a label set $L = \{l_1, \ldots, l_m\}$ via a linear head $g(h) = \sigma(Wh + b)$ with thresholding per label. For SCP, we perform token-level NER: each token t in a listing is assigned a type from a schema $\mathcal{Y}$ (e.g., `Brand`, `Model`, `Price`). Unless

stated otherwise, losses are standard binary cross-entropy for multi-label classification and cross-entropy for token tagging.

Multi-label classification. In multi-label text classification, one instance can carry several labels at once [Tsoumakas and Katakis, 2007]. Let $L = \{l_1, \ldots, l_n\}$ be the label set and A an ad. The goal is to predict a subset $L' \subseteq L$. A model f returns scores $P = \{p_1, \ldots, p_n\}$ with $p_i \in [0, 1]$ for each l_i, and a thresholding rule gives $\hat{L}' = \{l_i : p_i \geq \tau_i\}$. In our setting, this lets one CSA carry tags such as "massage" and "escort" at once. Class-imbalance is common; in low-label regimes, this pushes us toward data-efficient training and careful thresholding.

Named-entity recognition. NER extracts spans such as names, phones, places, and item descriptors; these fields are useful for triage and linking [Sundheim, 1995]. Unlike sentence classification, NER is token- or span-level. Sparse and noisy markup, code-mixing, and obfuscation raise the bar for low-label methods. Existing work has emphasized the challenge of obfuscation in this domain and explored advanced NER strategies to mitigate it [Perez et al., 2023].

Few-shot sentence-level fine-tuning. We use SetFit, a prompt-free few-shot method for sentence transformers [Tunstall et al., 2022]. SetFit follows two stages on top of Sentence Transformers Reimers and Gurevych, 2019: (1) *contrastive fine-tuning* on sentence pairs built from the small labeled set (same-label pairs as positives, different-label pairs as negatives) to learn a task-specific embedding space; (2) *lightweight classifier* training (e.g., logistic regression) over fixed embeddings. This yields fast training, low memory use, and stable performance when labels are scarce. Reports also show gains over full fine-tuning in small-data setups: accuracy improves by $\approx 0.06 - 1.7$ points and AUC by $4.2 - 6.6$ points in active-learning workflows [Schröder et al., 2023]. Beyond English-only cases, parameter-efficient and cross-lingual studies show that low-label methods can transfer well across languages and label sets [Adelani et al., 2023]. SetFit has also been tested in other few-shot text tasks (e.g., essay scoring, intent classification), where it is competitive with prompt-based methods under tight label budgets [Helmeczi et al., 2023]. Variants such as Anchoring fine-tuning inject semantic label signals into the embedding space to further help under few labels [Pauli et al., 2023].

Relation to prompt learning and sampling strategy. Prompt-learning frameworks (e.g., OpenPrompt) adapt pretrained models using task templates and verbalizers [Ding et al., 2022]. Prompting can be strong at very low k per class, but it adds design overhead and is sensitive to template choice. We opt for a prompt-free route to keep the pipeline simple and privacy-friendly (fewer labeled texts are needed). In imbalanced data, careful pair sampling for the contrastive stage and class-aware minibatches can improve stability, a pattern seen in recent small-data social-media studies as well [You et al., 2023].

Implications for our tasks. Sentence-level SetFit aligns well with post-level multi-label tagging, where the unit is a sentence or a short block. Token-level NER is harder because sentence embeddings do not expose span boundaries. This motivates our split design: (i) use SetFit for post-level triage, and (ii) assess limits at token level and note when span-specific heads or weak-label signals may be needed. The mix of methods above frames our approach: formal multi-label modeling [Tsoumakas and Katakis, 2007], classic NER goals [Sundheim, 1995], and a data-efficient SetFit pipeline on top of sentence transformers [Tunstall et al., 2022, Reimers and Gurevych, 2019], with recent evidence from active learning, cross-lingual work, and prompt-based alternatives [Schröder et al., 2023, Adelani et al., 2023, Helmeczi et al., 2023, Pauli et al., 2023, Ding et al., 2022, You et al., 2023].

Bridging sentence and token representations remains challenging; token-specific architectures and span-aware decoders are the de-facto standard for NER. Prompt-free few-shot methods such as SetFit complement, rather than replace, these token-level models. Contrastive learning has also been explored for NER to improve span separability, reinforcing our observation that sentence-level contrastive tuning alone is insufficient for fine-grained extraction.

```
Title: ❤️ nuru 🎀 gfe ☎️ 702--601--1555
Post: NEW staff working today. Private location near central strip.
Friendly, soft-spoken, and [SANITIZED] services for relaxed time.
Clean setting, same-day availability, respectful guests only.
We focus on comfort and stress relief; no drama, quick replies.
First-class session; leave happy and clear-minded.
Hours: 10am--10pm    Area: 1235 Laguna Ave, Las Vegas, NV 89169
Contact: ☎️ 702--601--1555    Note: Photos are representations.
```

Figure 2: Example of a CSA snippet (sanitized) associated with suspected HT activity. Possible labels: {Massage, Escort, Scam}.

3 Methodology

3.1 Classification for CSAs (HT risk)

Estimating HT risk in CSAs requires robust category definitions and a strategy that balances annotation efficiency with representational coverage. We approach this as a *multi-label post classification* task, where a single advertisement may simultaneously exhibit several risk-related attributes (e.g., both "massage" and "escort").

Category design. We first performed an exploratory analysis of the corpus to identify recurring lexical and semantic patterns, including references to services, payment schemes, and content formats. This yielded four coarse-grained but operationally meaningful categories:

1. *Massage*: ads explicitly offering massage services, often with implicit or explicit sexual references;
2. *Digital Entertainment*: ads offering digital media (photos, videos, live sessions) distributed via platforms such as OnlyFans or Snapchat;
3. *Escort*: ads indicating companionship or sexual services tied to events, locations, or social activities;
4. *Scam*: ads requesting upfront payment or providing vague service descriptions, consistent with fraudulent intent.

Our categorization is consistent with earlier NLP-based efforts to analyze sex-related advertisements for trafficking risk [Rodriguez Perez and Rivas, 2023]. For illustration, Figure 2 shows a sanitized HT ad snippet and its labels.

Annotation protocol. Each advertisement was reviewed and tagged with one or more of the above categories. To reduce annotation drift, annotators followed a fixed guideline, and ambiguous cases were resolved by majority consensus. This process produced a label distribution reflecting the real-world imbalance of CSAs associated with suspected HT activity.

Sentence-level representation. Because SetFit operates at the sentence level, we decomposed multi-sentence ads into individual units and aligned them with the ad's label set. We then explored three strategies for mapping sentence-level predictions back to the post level: (i) use the first sentence, (ii) use the last sentence, and (iii) compute the mean embedding across all sentences. This design tests whether early, late, or holistic textual cues carry more signal. Empirically, we find that averaging yields stronger context capture, as shown in Section 4. A simple variance-reduction argument formalizes this choice; see Appendix A (Theorem A.2 and Corollary A.3).

Rationale. This classification pipeline enables few-shot fine-tuning of sentence transformers in a high-stakes, low-resource setting. By explicitly modeling multi-label structure and testing sentence aggregation rules, we provide a reproducible framework for CSA risk detection that aligns with both the technical constraints of SetFit and the real-world complexity of online illicit advertising.

```
Title: Subaru full Titanium exhaust -- $800 (San Antonio)
Post: Full Tomei Titanium system off my 2012 STI; swapped to a
quieter setup. Need it gone ASAP.
About 3,000 miles of use; no dents, normal wear at flanges.
Price is $800 firm. Cash only. Local pickup preferred.
Text ☎ 2one0--232--9one99 for details; serious buyers please.
Tags: Subaru, STI, WRX, Impreza, Turbo, Exhaust, Titanium.
```

Brand	Subaru
Model	STI (2012)
Part Type	Exhaust (Tomei Titanium)
Price	$800
Phone	210--232--9199 (obfuscated in post)
Location	San Antonio

Figure 3: Suspected stolen-parts marketplace listing with entity extraction. Entities are shown as pairs below the snippet; note common obfuscation in the phone number.

3.2 NER for SCP marketplace listings

The extraction of entities from suspected stolen-parts listings within online marketplaces requires fine-grained entity extraction to recover salient details that may signal illicit activity. We formulate this task as an NER problem, where the goal is to segment an advertisement into spans and assign each span a type drawn from a predefined schema. In this domain, entities of interest include `Email`, `Location`, `Name`, `Phone Number`, `Date`, `Part Type`, `Brand`, `Model`, and `Price`. These fields support linking posts, characterizing sellers, and mapping transactional patterns.

Annotation protocol. To construct supervision signals, annotators manually reviewed advertisements and highlighted token spans corresponding to the schema above. Given the heterogeneous nature of the text, often informal, code-mixed, and noisy, annotators were trained with domain-specific guidelines to ensure consistency. Ambiguities were adjudicated by consensus, yielding a modest but reliable labeled set. This process reflects real-world constraints, where sensitive data access and privacy concerns limit large-scale annotation.

Modeling considerations. Unlike post-level classification, NER operates at the token level. This presents challenges for SetFit, which is designed for sentence-level representations. To adapt, we segmented ads into sentences or short fragments and treated each fragment as a candidate unit for token-level classification. Tokens were aligned with their fragment-level embeddings and trained against the entity labels. While this setup enables few-shot fine-tuning, it exposes a structural mismatch: embeddings optimized for semantic similarity at the sentence level must now capture sub-sentence, entity-specific distinctions. Figure 3 shows a suspected stolen-parts listing with extracted entities. Concretely, for each token we used the embedding of its enclosing sentence, $h_i = E(s_i)$, as the token feature (i.e., every token in s_i shares h_i), which we then fed to a token classifier; this sentence-to-token mapping is the locus of the observed mismatch.

Motivation. By applying SetFit to this task, we aim to probe its limits in fine-grained extraction and assess how far sentence-level embeddings can be stretched to token-level recognition under low-resource conditions. The results, discussed in Section 4, highlight both the promise and the shortcomings of this approach, motivating future work on hybrid architectures that combine few-shot contrastive training with span-sensitive token classifiers.

3.3 SetFit processing

Both the classification and NER tasks described above are constrained by the scarcity of high-quality annotated data and the difficulty of producing consistent labels in sensitive domains. Manual annotation requires familiarity with trafficking-related terminology and domain-specific jargon in automotive parts, as well as careful handling of noisy and obfuscated text. These factors make conventional supervised training impractical at scale. To address

this, we adopt `SetFit`, a few-shot framework designed to fine-tune sentence transformers without prompts [Tunstall et al., 2022].

Model adaptation. SetFit operates in two stages: (i) contrastive fine-tuning of a sentence transformer using pairs of labeled examples, and (ii) training a lightweight classifier (e.g., logistic regression) on the resulting embeddings. This design yields strong performance with only a handful of labeled samples while avoiding the overhead of full parameter updates. For our tasks, we decompose each advertisement into sentences, align these sentences with the ad's labels, and then construct positive and negative pairs for contrastive training. This enables SetFit to learn task-specific semantic spaces from limited supervision. Figure 1 summarizes the pipeline from raw ad text to predictions and marks the stages used for post-level aggregation and token-level adaptation.

Post-level aggregation. Because ads typically contain multiple sentences, we evaluate three aggregation strategies for deriving a post-level prediction from sentence-level embeddings: (a) first-sentence representation, (b) last-sentence representation, and (c) mean pooling over all sentences. These strategies capture different hypotheses about where the most salient cues for HT risk in CSAs may appear. The averaged embedding strategy emerges as the strongest in our experiments (see Section 4).

Token-level adaptation. Applying SetFit to NER requires extending its sentence-level embeddings to finer granularity. We segment ads into short fragments and align token spans with the corresponding embeddings, training a classifier to predict entity labels such as `Phone Number`, `Brand`, or `Price`. This repurposing exposes a structural limitation: embeddings tuned for semantic similarity at the sentence level are not naturally optimized for token-level recognition. Nonetheless, this setup allows us to quantify the degree to which SetFit can be stretched toward fine-grained extraction in low-resource conditions. This limitation aligns with other evidence that generic embeddings often require task-specific adaptation for trafficking risk prediction [Arguelles Terron et al., 2023].

Summary. This pipeline leverages SetFit's efficiency to build models under severe data scarcity, while also stress-testing its limits across both sentence-level and token-level tasks. By systematically exploring aggregation strategies and evaluating its generalization to NER, we provide insights into the adaptability of SetFit for real-world, high-stakes applications involving illicit online advertising.

4 Experimental Results

We evaluate the effectiveness of SetFit in two tasks: multi-label classification of CSAs (HT risk) and NER in SCP listings. All experiments use the `all-roberta-large-v1` backbone from the Sentence-Transformers library [Reimers and Gurevych, 2019], which provides sentence-level RoBERTa models. Unless otherwise stated, we report macro-averaged F1 scores. For CSA classification we used 232 labeled posts across m labels; for SCP NER we used 354 (SetFit) and 1000 (GPT-2) labeled listings, respectively.

Human trafficking classification. We first consider the post-level classification task. Since CSAs often contain multiple sentences, we tested three sentence-aggregation strategies when forming post embeddings: (i) *first-sentence* only, (ii) *last-sentence* only, and (iii) *mean aggregation* across all sentences. With only 232 labeled ads, SetFit with mean aggregation achieved the strongest performance, reaching an F1 score of **0.783**. In contrast, first- and last-sentence heuristics yielded F1 scores of 0.627 and 0.629, respectively. These results highlight the importance of capturing the full context of an ad rather than relying on a single sentence cue.

To contextualize these findings, we trained a GPT-2 baseline [Radford et al., 2019] with 400 labeled examples, nearly twice the training data used by SetFit, and a parameter count roughly five times larger (1.5B vs. 355M). GPT-2 attained an F1 of 0.703, confirming that SetFit not only requires fewer annotations but also outperforms a much larger generative baseline in this classification setting. This underscores SetFit's suitability for low-resource,

high-stakes domains where labeled data is expensive or sensitive. These outcomes confirm the annotation and dataset challenges highlighted in other recent work on HT corpora [Rivas et al., 2024].

Stolen car parts NER. We next assess SetFit on token-level entity extraction. Here, the model is tasked with identifying entities such as `Brand`, `Model`, `Price`, and `Phone Number`. In contrast to classification, this task demands fine-grained token-level distinctions that sentence-level embeddings are not optimized to capture. Consistent with this mismatch, SetFit achieved only an F1 score of 0.242, despite being trained on 354 annotated examples.

For comparison, GPT-2 trained on 1000 labeled samples achieved a higher F1 of **0.337**. Although still modest, the gap illustrates SetFit's limitations in extending from sentence-level similarity learning to fine-grained token labeling. This performance degradation emphasizes the need for hybrid strategies, e.g., span-based encoders, CRF decoding, or weak supervision, that better exploit structure at the token level.

Qualitatively, the most common failure modes were: (i) missed short spans when semantics were carried by obfuscated tokens (e.g., digit/letter mixes), and (ii) boundary drift where prices or brands were only partially captured in code-mixed phrases.

Findings. Across both tasks, results reveal a clear dichotomy: SetFit is highly effective for few-shot post classification, outperforming a much larger GPT-2 baseline while consuming fewer labels, but struggles when transferred to NER. These findings suggest that while SetFit is well-suited for document- or sentence-level classification in illicit advertising detection, token-level entity extraction requires specialized architectures. Table 1 provides a consolidated overview of these results.

Table 1: Performance evaluation of SetFit and GPT-2 models in CSA (HT Risk) post classification and SCP NER on marketplace listings. SetFit (F), (L), and (A) denote classification by first sentence, last sentence, and average over all sentences, respectively.

Task	Method	Train	Params	F1
CSA (HT Risk) Class	SetFit (F)	232	355M	0.627
	SetFit (L)	232	355M	0.629
	SetFit (A)	232	355M	**0.783**
	GPT-2	400	1.5B	0.703
SCP NER	SetFit	354	355M	0.242
	GPT-2	1000	1.5B	**0.337**

5 Conclusions

This work investigated the application of SetFit, a prompt-free few-shot learning framework, to the detection of illicit online advertisements. Our study focused on two high-stakes domains: CSAs (HT risk) and SCP marketplace listings. The experiments yield three main findings.

First, SetFit proves highly effective for multi-label classification of CSAs (HT risk). Despite training on fewer than 250 labeled examples, the averaging strategy for sentence aggregation achieved an F1 of **0.783**, outperforming a GPT-2 baseline trained on nearly twice as many examples and with a parameter count five times larger. This result highlights the suitability of SetFit for low-resource, socially critical domains, where annotation is costly, sensitive, or privacy-limited.

Second, our results expose the limits of SetFit when extended to token-level NER. On the SCP dataset of marketplace listings, SetFit achieved only 0.242 F1, compared to 0.337 for GPT-2 trained with more supervision. This gap underscores the structural mismatch between sentence-level embeddings optimized for semantic similarity and the fine-grained sequence modeling required for entity extraction. The findings suggest that SetFit's current design

is well-matched to post-level classification but requires augmentation with span-sensitive architectures or hybrid approaches for token-level tasks.

Finally, the methodological choices examined here, particularly the aggregation strategies for multi-sentence posts, provide concrete guidance for applying few-shot models in real-world safety-critical contexts. Our analysis shows that simple, reproducible design decisions can significantly impact performance and are therefore critical to document in applied NLP research. These findings underscore the need for interdisciplinary approaches and visualization tools to support a broader understanding of trafficking and related illicit networks [Gortney et al., 2025, Giddens et al., 2023].

Our NER comparison is limited to a GPT-2 baseline; standard token heads (e.g., RoBERTa-large with a CRF layer) and few-shot prototype-based NER are natural comparators we defer to future work. We also view SetFit embeddings plus span-sensitive decoders (e.g., CRF/pointer) as a promising hybrid for SCP.

The limitations observed in NER highlight several promising avenues: integrating SetFit embeddings with span-based or CRF decoders, leveraging weak supervision to expand token-level labels, and combining contrastive few-shot training with prompt-based techniques. Exploring these directions may yield models that retain the data efficiency of SetFit while addressing its shortcomings in fine-grained extraction. Additionally, multilingual extensions are essential to broaden applicability to the diverse linguistic landscape of online illicit markets.

Overall, our results demonstrate that SetFit offers a practical and efficient framework for detecting harmful online content when annotation budgets are constrained. By advancing few-shot methodologies in socially impactful domains, this study contributes both empirical evidence and methodological insights to the broader effort of building robust, responsible AI tools for public safety. We expect the observed split, strong post-level few-shot performance vs. weak token-level transfer, to generalize to other low-resource, fine-grained tasks (e.g., slot filling or clinical entity spans).

Acknowledgments and Disclosure of Funding

The authors thank the Rivas.AI Lab (`https://lab.rivas.ai`) for the support and helpful feedback throughout this project. Part of this work was funded by the National Science Foundation under grants CNS-2136961 and CNS-2210091.

References

D. Adelani, M. Masiak, I. Azime, J. Alabi, A. Tonja, C. Mwase, O. Ogundepo, B. Dossou, A. Oladipo, D. Nixdorf, C. Emezue, S. Al-Azzawi, B. Sibanda, D. David, L. Ndolela, J. Mukiibi, T. Ajayi, T. Moteu, B. Odhiambo, A. Owodunni, N. Obiefuna, M. Mohamed, S. Muhammad, T. Ababu, S. Salahudeen, M. Yigezu, T. Gwadabe, I. Abdulmumin, M. Taye, O. Awoyomi, I. Shode, T. Adelani, H. Abdulganiyu, A. Omotayo, A. Adeeko, A. Afolabi, A. Aremu, O. Samuel, C. Siro, W. Kimotho, O. Ogbu, C. Mbonu, C. Chukwuneke, S. Fanijo, J. Ojo, O. Awosan, T. Kebede, T. Sakayo, P. Nyatsine, F. Sidume, O. Yousuf, M. Oduwole, K. Tshinu, U. Kimanuka, T. Diko, S. Nxakama, S. Nigusse, A. Johar, S. Mohamed, F. Hassan, M. Mehamed, E. Ngabire, J. Jules, I. Ssenkungu, and P. Stenetorp. Masakhanews: news topic classification for african languages. 2023. doi: 10.18653/v1/2023.ijcnlp-main.10.

Sara Aniello and Stefano Caneppele. Selling stolen goods on the online markets: An explorative study. *Global Crime*, 19(1):42–62, 2018.

Ana Paula Arguelles Terron, Jorge Yero Salazar, Pablo Rivas, Ernesto Quevedo Caballero, and Alejandro Rodriguez Perez. Task-specific or task-agnostic? a statistical inquiry into BERT for human trafficking risk prediction. In *Neural Information Processing Systems Conference: LatinX in AI (LXAI) Research Workshop*, New Orleans, Louisiana, 2023. doi: 10.52591/lxai202312106. URL `https://doi.org/10.52591/lxai202312106`. Poster presentation.

R. Bensoltane and T. Zaki. Enhancing arabic offensive language detection with bert-bigru model. *Bulletin of Electrical Engineering and Informatics*, 13:1351–1361, 2024. doi: 10.11591/eei.v13i2.6530.

N. Ding, S. Hu, W. Zhao, Y. Chen, Z. Liu, H. Zheng, and M. Sun. Openprompt: an open-source framework for prompt-learning. 2022. doi: 10.18653/v1/2022.acl-demo.10.

Laurie Giddens, Stacie Petter, Gisela Bichler, Pablo Rivas, Michael H. Fullilove, and Tomas Cerny. Navigating an interdisciplinary approach to cybercrime research. In *Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS-56)*, Online, 2023. URL `https://aisel.aisnet.org/hicss-56/in/cybercrime/2`.

Mia Gortney, Garret Parker, Patrick Harris, Austin Huizinga, Alejandro Rodriguez Perez, Ernesto Quevedo Caballero, Kor Sooksatra, Tomas Cerny, Pablo Rivas, Bikram Khanal, Maisha Binte Rashid, and Jie Ren. Visualizing human trafficking and criminal networks: A systematic mapping study. In *Proceedings of the 24th International Conference on Information & Knowledge Engineering (IKE'25), The 2025 World Congress in Computer Science, Computer Engineering, and Applied Computing (CSCE'25)*, Las Vegas, NV, USA, July 2025. July 21–24, 2025.

R. Helmeczi, S. Yıldırım, M. Çevik, and S. Lee. Few shot learning approaches to essay scoring. 2023. doi: 10.21428/594757db.8702fa2f.

V. Koreddi, N. Manisha, S. Kaif, and Y. Kumar. Multilingual ai system for detecting offensive content across text, audio, and visual media. *Engineering Research Express*, 7: 015216, 2025. doi: 10.1088/2631-8695/ada72a.

Nicholas D Kristof. How pimps use the web to sell girls. *The New York Times*, 25, 2012.

G. Mou, Y. Yue, K. Lee, and Z. Zhang. Wildlife product trading in online social networks: a case study on ivory-related product sales promotion posts. *Proceedings of the International Aaai Conference on Web and Social Media*, 18:1096–1109, 2024. doi: 10.1609/icwsm.v18i1.31375.

A. Pauli, L. Derczynski, and I. Assent. Anchoring fine-tuning of sentence transformer with semantic label information for efficient truly few-shot classification. 2023. doi: 10.18653/v1/2023.emnlp-main.692.

Alejandro Rodriguez Perez, Pablo Rivas, Javier Turek, Korn Sooksatra, Ernesto Quevedo, Gisela Bichler, Tomas Cerny, Laurie Giddens, and Stacie Petter. Decoding the obfuscated: Advanced ner techniques on commercial sex advertisement data. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 144–151, 2023. doi: 10.1109/CSCI62032.2023.00029.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Pablo Rivas, Tomas Cerny, Alejandro Rodriguez Perez, Javier Turek, Laurie Giddens, Gisela Bichler, and Stacie Petter. On the challenges of creating datasets for analyzing commercial sex advertisements to assess human trafficking risk and organized activity. *arXiv*, May 2024. doi: 10.48550/arXiv.2405.13348. URL `https://arxiv.org/abs/2405.13348`. LXAI Workshop at NAACL 2024.

Alejandro Rodriguez Perez and Pablo Rivas. Combatting human trafficking in the cyberspace: A natural language processing-based methodology to analyze the language in online advertisements. *arXiv*, November 2023. doi: 10.48550/arXiv.2311.13118. URL `https://arxiv.org/abs/2311.13118`.

C. Schröder, L. Müller, A. Niekler, and M. Potthast. Small-text: active learning for text classification in python. pages 84–95, 2023. doi: 10.18653/v1/2023.eacl-demo.11.

Beth M Sundheim. Overview of results of the muc-6 evaluation. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, 1995.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*, 2022. URL `https://arxiv.org/abs/2209.11055`.

W. You, H. Yang, Z. Ding, X. Zhou, Y. Zhou, L. Ma, and Z. Hou. Trajectories of and spatial variations in hpv vaccine discussions on weibo, 2018-2023: a deep learning analysis. 2023. doi: 10.1101/2023.12.07.23299667.

L. Yuan, H. Jiang, H. Shen, L. Shi, and N. Cheng. Sustainable development of information dissemination: a review of current fake news detection research and practice. *Systems*, 11: 458, 2023. doi: 10.3390/systems11090458.

A Appendix: Why Mean-over-Sentences Improves Post-level Scoring

Setup. Each post consists of n sentences with embeddings $s_1, \ldots, s_n \in \mathbb{R}^d$ produced by the SetFit-tuned encoder. Let $y \in \{0, 1\}$ be the post label for a given class (one-vs-rest). Assume a standard linear score $g(x) = w^\top x + b$ used by the lightweight classifier. We compare three post-level representations: FIRST s_1, LAST s_n, and MEAN $\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i$.

Assumptions. (A1) Conditional on y, sentence embeddings are unbiased around a class mean: $\mathbb{E}[s_i \mid y] = \mu_y$. (A2) Sentence-wise noise is zero-mean with covariance Σ_y, independent across i: $s_i = \mu_y + \varepsilon_i$, $\mathbb{E}[\varepsilon_i \mid y] = 0$, $\mathrm{Cov}(\varepsilon_i \mid y) = \Sigma_y$. (A3) The linear head g is fixed (e.g., logistic regression after SetFit).

These assumptions match the case where the SetFit contrastive stage pulls same-class sentences toward a prototype while leaving sentence-level noise (formatting, slang, markup) as additive variation.

Lemma A.1 (Variance reduction for the mean)**.** *Under (A1)-(A2),* $\mathbb{E}[\bar{s} \mid y] = \mu_y$ *and* $\mathrm{Cov}(\bar{s} \mid y) = \Sigma_y/n$*. Hence the score variance satisfies*

$$\mathrm{Var}[g(\bar{s}) \mid y] = w^\top \Sigma_y w \;/\; n,$$

while for any single sentence (e.g., FIRST *or* LAST*),*

$$\mathrm{Var}[g(s_1) \mid y] = \mathrm{Var}[g(s_n) \mid y] = w^\top \Sigma_y w.$$

Proof. Linearity of expectation and independence across sentences give the mean and covariance of $\bar{s}$; projecting through w yields the scalar variance formulas. □

Theorem A.2 (Mean aggregation improves margin SNR)**.** *Let* $\Delta\mu = w^\top(\mu_1 - \mu_0)$*. Under (A1)-(A3), the class-conditional score means satisfy* $\mathbb{E}[g(\cdot) \mid y] = w^\top \mu_y + b$*. The signal-to-noise ratio (SNR) of the* MEAN *score,*

$$\mathrm{SNR}_{MEAN} = \frac{|\Delta\mu|}{\sqrt{w^\top \Sigma_0 w/n + w^\top \Sigma_1 w/n}},$$

is a factor $\sqrt{n}$ *larger than the SNR for any single-sentence score (*FIRST/LAST*).*

Proof. Immediate from Lemma A.1 applied to both classes and the definition of SNR for a difference of Gaussians (or sub-Gaussian) with independent noise terms. □

Corollary A.3 (AUC improvement for Gaussian scores)**.** *If* $(g(\cdot) \mid y = 0)$ *and* $(g(\cdot) \mid y = 1)$ *are Gaussian, then*

$$\mathrm{AUC} \;=\; \Phi\!\left(\frac{\Delta\mu}{\sqrt{\mathrm{Var}(g(\cdot) \mid y=0) + \mathrm{Var}(g(\cdot) \mid y=1)}}\right).$$

By Lemma A.1, $\mathrm{AUC}_{\mathit{MEAN}} \geq \mathrm{AUC}_{\mathit{FIRST}} = \mathrm{AUC}_{\mathit{LAST}}$*, with strict inequality when* $w^\top \Sigma_y w > 0$*.*

Proof. For two normal score distributions with equal priors, AUC is a monotone function of the standardized mean gap; reducing both class variances by $1/n$ increases the argument of $\Phi(\cdot)$. □

Implications. (i) **Why mean wins.** The mean has the same expected margin as any single sentence but smaller variance (by $1/n$), improving ROC metrics and typically F1 after proper thresholding. (ii) **When first/last can tie.** Only in degenerate cases where $w^\top \Sigma_y w = 0$ (no sentence-level noise along w) or when the informative content appears in exactly one sentence and others are pure noise with extreme outliers that bias $\bar{s}$. (iii) **Weighted pooling.** If sentences have unequal noise levels, the BLUE solution gives weights $\propto \Sigma_y^{-1}$, i.e., inverse-variance pooling. If such estimates are unavailable, uniform mean is minimax-stable. (iv) **Beyond linear heads.** For smooth $g(\cdot)$, a first-order Taylor expansion around μ_y yields the same $1/n$ variance reduction leading to improved separation; details are standard.

Connection to SetFit contrastive training. With positive pairs formed from same-label sentences, contrastive fine-tuning encourages within-class concentration around μ_y and between-class separation along w. Under this lens, $\bar{s}$ is an efficient estimator of the latent post-level prototype, explaining the empirical gains we observe for mean pooling in CSA classification.