SWEET - Weakly Supervised Person Name Extraction for Fighting Human Trafficking

Javin Liu*, Hao Yu*, Vidya Sujaya*, Pratheeksha Nair, Kellin Pelrine, Reihaneh Rabbany

Infrared AI for Humanity Mila · Mila · McGill

Github Link: https://github.com/ComplexData-MILA/SWEET

Problem Formulation

(WHAT) How can we extract person names from escort advertisements, where the text

1. is noisy
2. includes sensitive language
3. contains private information
4. is lacking labelled data

(WHY) Application:

1. help clarify information in online escort ads
2. used to pinpoint possible human trafficking (HT)

(HOW) **SWEET**: A weak supervision pipeline that

a. combines fine-tuned language models and
b. antirules to extract person names
c. without the need for task-specific training labels.
d. no human labeling is required

(HOW Performance) Compared to the previous supervised SOTA method for this task, SWEET has:

- 10% higher F1 score on HT domain datasets
- 70% higher F1 score on benchmark datasets, better generalization

Dataset

(**A**) Evaluation, and (**B**) LM fine-tuning.

Dataset Name	Purpose	Train	Test	Batch Size
HTName	A	-	995	-
HTUnsup	B	6,160	-	32
HTGen	A & B	9424	818	32
FewNERD-L1	B	131,767	18,824	32
WikiNER-en	B	129,907	14,435	128
CoNLL2003	A & B	14,041	3,453	128
WNUT2017	A & B	3,394	1,287	128

[Private Datasets]

HTName: in-domain evaluation dataset

HTUnsup: in-domain fine-tuning dataset gathered from private escort websites, labelled by ChatGPT (performance reported below)

F1	Precision	Recall
.901	.894	.908

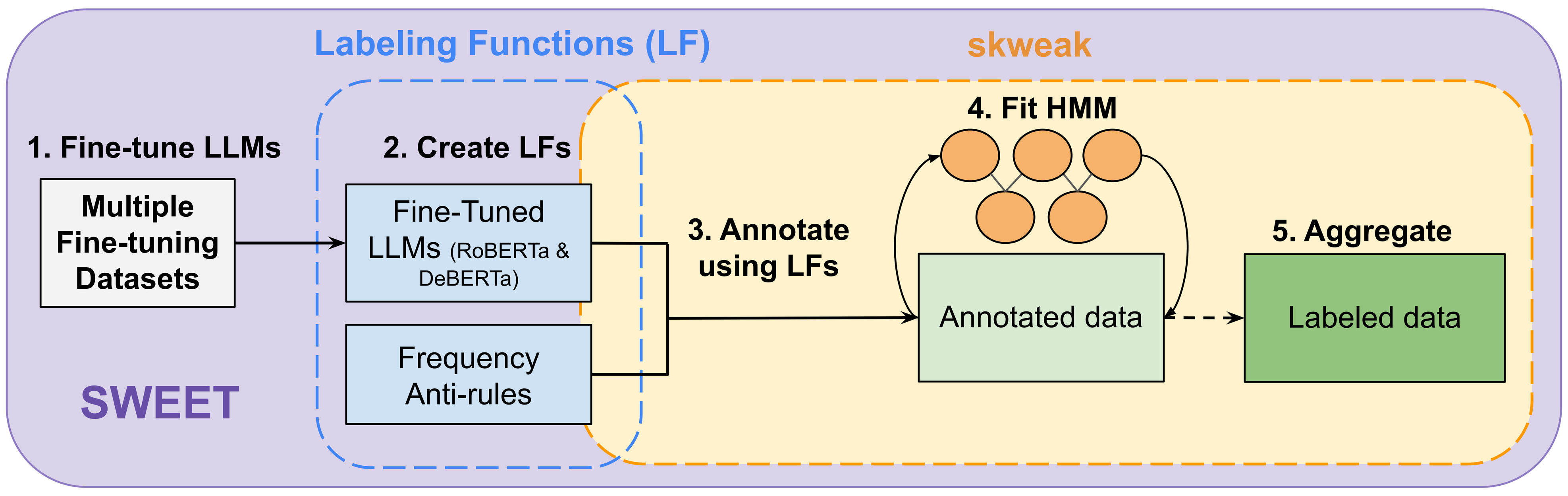
ChatGPT prompt:

I want you to act as a natural and no-bias labler, extract human's name and location or address and social media link or tag in the format 'Names: \nLocations: \nSocial: '. If exists multiple entities, separated by |. If not exists, say N. Your words should extract from the given text, can't add/modify any other words. As shorter as possible, remember don't include phone number. For one name, should be less than 3 words.

[Open-source Datasets]

From HuggingFace and other public sources, we used the training sets of FewNERD-L1, WikiNER-en, CoNLL2003, and WNUT2017 to fine-tune our language models, and the test split of the latter two for evaluation.

Method



1. **Fine-tune** *DeBERTa V3 and RoBERTa* models on our type B datasets for NER task.
2. **Create** *labeling functions* (LFs). The fine-tuned models are LFs that annotate words as *"PERSON_NAME"*, while antirules counter possible noise by annotating words as *"NOT_NAME"*.
3. **Annotate using LFs.** We vary SWEET by using LF subsets, specified in the results table.
4. **Fit** *a hidden markov model (HMM)* on the annotated dataset following the skweak approach (Lison et al., 2021).
5. **Aggregate** *all annotations* by applying the fitted HMM on the same dataset used to fit it. The HMM's output is our final label.

Why HMM?

a. A word may have multiple LFs determining it as a *name* or *not*. An HMM is used as an aggregator to resolve possible conflicts.
b. Its states and observations correspond to the true labels and LF outputs respectively.
c. Initial parameters are calculated using majority vote results, and later estimated using the Baum-Welch algorithm.
d. Each LF has a weight tempered in the process, decreasing based on redundancy (using recall with other LFs as a measure).

Example

	HTName	CoNLL2003
Input Text	HI MIA HERE FIRST TIME IN THIS CITY,WOULD LIKE TO MEET NICE GUYS... COME MEET ME TO HAVE UNFORGETTHABLE TIME TOGETHER...NEVER RUSH OPEN - MINDED MENU CALL TEXT... 123456789 EGLINTON AVE E SCARBOROUGH	China controlled most of the match and saw several chances missed until the 78th minute when Uzbek striker Igor Shkvyrin took advantage of a misdirected defensive header to lob the ball over the advancing Chinese keeper and into an empty net.
spaCy baseline	'Eglinton'	'Striker Igor Shkvyrin'
NEAT (previous SOTA) baseline	'MIA'	-
SWEET	'MIA'	'Igor', 'Shkvyrin'
Ground Truth	'MIA'	'Igor', 'Shkvyrin'

Results

Method	HTName F1	HTName Prec	HTName Rec	HTGen F1	HTGen Prec	HTGen Rec	CoNLL2003 F1	CoNLL2003 Prec	CoNLL2003 Rec	WNUT2017 F1	WNUT2017 Prec	WNUT2017 Rec
spaCy (Honnibal et al., 2020)	$.27 \pm .03$	$.18 \pm .02$	$.51 \pm .02$	$.47 \pm .04$	$.50 \pm .03$	$.43 \pm .03$	$.64 \pm .04$	$.66 \pm .04$	$.55 \pm .04$	$.21 \pm .07$	$.14 \pm .06$	$.44 \pm .06$
TwitterNER (Mishra and Diesner, 2016)	$.56 \pm .04$	$75. \pm .04$	$52. \pm .04$	$.70 \pm .02$	$.70 \pm .03$	$.67 \pm .03$	$.68 \pm .05$	$.91 \pm .05$	$.55 \pm .05$	$.61 \pm .09$	$.84 \pm .10$	$.57 \pm .10$
LUKE (Yamada et al., 2020)	$.63 \pm .03$	$.85 \pm .04$	$.51 \pm .04$	$.68 \pm .03$	$.84 \pm .02$	$.56 \pm .02$	$.31 \pm .11$	$.89 \pm .09$	$.19 \pm .09$	$.55 \pm .07$	$.67 \pm .05$	$.44 \pm .05$
ELMo (Peters et al., 2018)	$.51 \pm .02$	$.56 \pm .02$	$.46 \pm .02$	$.69 \pm .05$	$.61 \pm .06$	$.74 \pm .06$	$.96 \pm .02$	$.95 \pm .02$	$.99 \pm .02$	$.59 \pm .15$	$.72 \pm .18$	$.37 \pm .18$
Flair (Akbik et al., 2019)	$.45 \pm .02$	$.73 \pm .02$	$.32 \pm .02$	$.63 \pm .04$	$.83 \pm .04$	$.49 \pm .04$	$.98 \pm .02$	$.97 \pm .02$	$1.0 \pm .02$	$.60 \pm .15$	$.79 \pm .18$	$.34 \pm .18$
NEAT (Original) (Li et al., 2022)	$.78 \pm .04$	$.83 \pm .05$	$.74 \pm .03$	$.71 \pm .01$	$.63 \pm .02$	$.79 \pm .03$	$.17 \pm .07$	$.43 \pm .05$	$.07 \pm .05$	$.22 \pm .06$	$.47 \pm .04$	$.16 \pm .04$
NEAT (Weakly Supervised)	$.79 \pm .02$	$.80 \pm .02$	$.77 \pm .02$	$.71 \pm .01$	$.64 \pm .02$	$.78 \pm .02$	$.16 \pm .07$	$.42 \pm .05$	$.07 \pm .05$	$.22 \pm .06$	$.47 \pm .04$	$.16 \pm .04$
Majority vote	$.73 \pm .02$	$.59 \pm .01$	$.95 \pm .01$	$.74 \pm .02$	$.65 \pm .03$	$.85 \pm .03$	$.83 \pm .06$	$.74 \pm .02$	$.98 \pm .02$	$.65 \pm .04$	$.53 \pm .06$	$.90 \pm .06$
SWEET – *Domain Data*	$.88 \pm .01$	$.85 \pm .01$	$.92 \pm .01$	$.75 \pm .02$	$.71 \pm .03$	$.78 \pm .03$	$.86 \pm .05$	$.79 \pm .02$	$.97 \pm .02$	$.69 \pm .06$	$.61 \pm .06$	$.82 \pm .06$
SWEET	$.87 \pm .01$	$.83 \pm .01$	$.92 \pm .01$	$.81 \pm .02$	$.76 \pm .03$	$.84 \pm .03$	$.86 \pm .05$	$.79 \pm .03$	$.98 \pm .03$	$.68 \pm .04$	$.58 \pm .07$	$.83 \pm .07$

Model	Fine-tuning Dataset	F1	Precision	Recall
DeBERTa-v3-base	HTUnsup	$.67 \pm .03$	$.71 \pm .02$	$.62 \pm .02$
	HTGen	$.68 \pm .01$	$.71 \pm .02$	$.67 \pm .02$
	CoNLL2003	$.67 \pm .02$	$.67 \pm .03$	$.69 \pm .03$
	Few-NERD-L1	$.57 \pm .03$	$.80 \pm .03$	$.43 \pm .03$
	WikiNER-en	$.52 \pm .01$	$.48 \pm .02$	$.54 \pm .02$
	WNUT2017	$.70 \pm .02$	$.71 \pm .02$	$.72 \pm .02$
RoBERTa-base	HTUnsup	$.82 \pm .02$	$.84 \pm .03$	$.83 \pm .03$
	HTGen	$.72 \pm .02$	$.81 \pm .02$	$.66 \pm .02$
	CoNLL2003	$.72 \pm .02$	$.68 \pm .03$	$.77 \pm .03$
	Few-NERD-L1	$.68 \pm .02$	$.81 \pm .02$	$.59 \pm .02$
	WikiNER-en	$.49 \pm .03$	$.43 \pm .03$	$.56 \pm .03$
	WNUT2017	$.68 \pm .02$	$.73 \pm .03$	$.66 \pm .03$

Conclusion

- **SWEET** obtains SOTA on HTName of 0.87 F1
- **SWEET** generalizes better to benchmark datasets
- **SWEET** maintains/improves performance on removing domain data LFs
- **SWEET** does not require any human annotators
- **SWEET** easy to expand to other domains with more LFs

Footnotes

*These authors contributed equally to this work

References

1. Lison, Pierre, Jeremy Barnes, and Aliaksandr Hubin. skweak: Weak supervision made easy for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics.
2. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A., & others. (2020). spaCy: Industrial-strength natural language processing in Python. Zenodo, Honolulu, HI, USA.
3. Li, Y., Nair, P., Pelrine, K., & Rabbany, R. (2022). Extracting Person Names from User Generated Text: Named-Entity Recognition for Combating Human Trafficking. In Findings of the Association for Computational Linguistics: ACL 2022 (pp. 2854-2868).