Online Sexism Detection

Tanya Roosta, Ph.D.
Senior Applied Science Manager at Alexa AI, Amazon

This presentation is based on the following paper that is under review at SemEval-2023 Task 10

"***Alexa at SemEval-2023 Task 10: Explainable Detection of Online Sexism***"

Authors:
Weston Feely∗, Prabhakar Gupta∗, Manas Mohanty,
Timothy Chon, Tuhin Kundu,
Vijit Singh, Sandeep Atluri, Tanya Roosta, Viviane Ghaderi,
Peter Schulam, Heba Elfardy

* Equal contribution

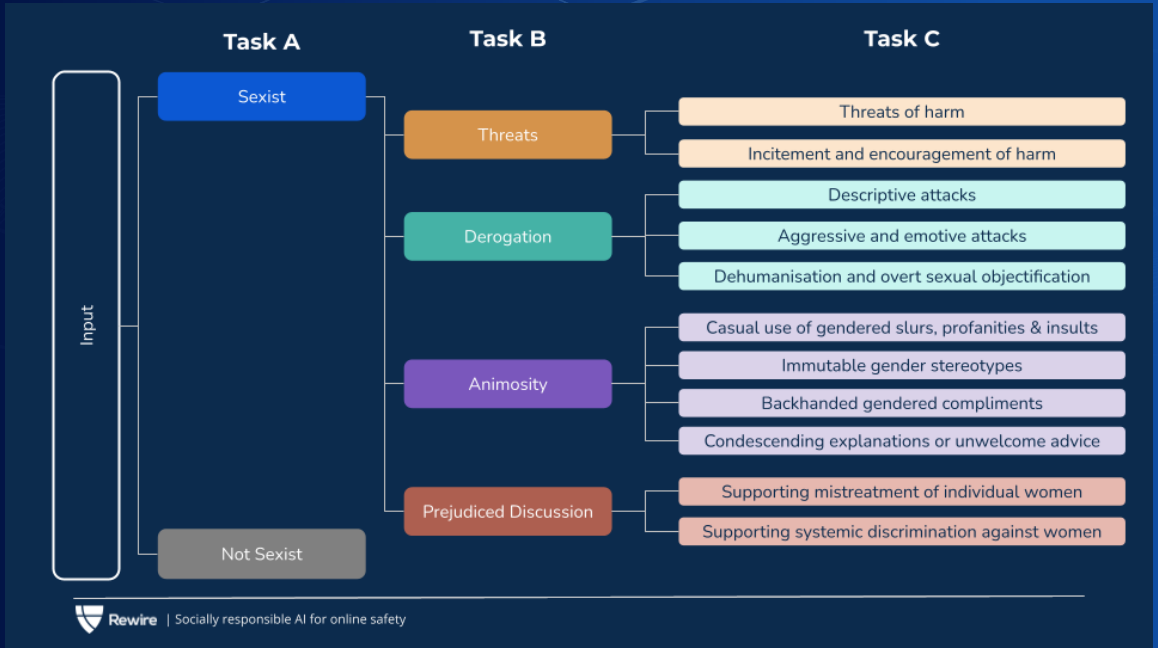
What is SemEval?

- **SemEval** is a series of international natural language processing (NLP) research workshops whose mission is to:
  - Advance the current state of the art in semantic analysis
  - Help create high-quality annotated datasets in a range of increasingly challenging problems in natural language semantics
  - Each year's workshop features a collection of shared tasks in which computational semantic analysis systems designed by different teams are presented and compared.

SemEval Task 10

“Explainable Detection of Online Sexism (EDOS)”

This task supports the development of English-language models for sexism detection that are more accurate as well as *explainable*, with fine-grained classifications for sexist content from Gab and Reddit.



Online Sexism

Sexism*
Any abuse or negative sentiment that is directed towards women based on their gender, or based on the gender combined with one or more other identity attributes (e.g. Black women, Muslim women, Trans women)

Sexism online is a growing concern especially on social media apps
It can cause mental harm to individuals who are targeted, makes online interactions unwelcoming, and spreads injustice

Explainable Detection
Many automated tools that are widely used can find sexist content at scale, but they only give high-level categories. Flagging what is sexist and explaining why improves interpretability, trust and understanding of the decisions made by automated tools*

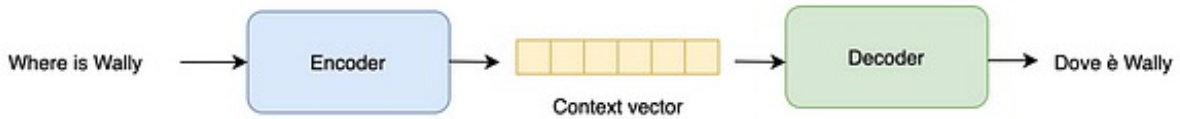
*: https://codalab.lisn.upsaclay.fr/competitions/7124

Machine Learning Solutions

Goal: purely automatic, machine learning approach to the problem of sexism detection and classification in social media posts

The current dominant approach to toxic language detection is to fine-tune a sequence to sequence (encoder/decoder) transformer model

What is a sequence to sequence model?



The architecture of the model inside the encoder and decoder components can have a variety of forms

Transformer architecture is the state-of-the-art used in all sequence to sequence models (e.g. BERT, RoBERTa)

Image: https://medium.com/nlplanet/two-minutes-nlp-visualizing-seq2seq-models-with-attention-10020e233b6c

Our Modeling Solution

Ensemble methods are very common in building predictive models, random forests being one of the best known examples

A recent thread of research has analyzed the accuracy of ensemble of DNN models and showed improvements*

We used different ensembling of the following models for each task.

- BERT base-uncased
- ELECTRA base and large
- XLM-RoBERTa large

The models vary in the selection of: base models, hyper-parameter configurations, and the data used to fine-tune each model

* Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time

Fine-Tuning of Models

We fine-tune models for two sets of ensemble:

- Task A
- Task C
  - For task B, we use mapping from the fine-grained labels of task C to coarse-grained labels of task B

For task A, we used cross-entropy loss, whereas for multi-class tasks B and C, we used a shared loss function:

- Composite loss computed as the sum of (1) the standard cross entropy (CE) loss on all 11 Task C outputs, and (2) the cross entropy of the Task C outputs projected onto the 4 outputs for Task B

$$\mathcal{L}_C = \beta \cdot CE(Y_B, \hat{Y}_B) + CE(Y_C, \hat{Y}_C)$$

In addition, we use in-context learning to prompt GPT-3, and used the results in our ensemble.

Datasets

- SemEval Task 10 provided the corpus of gold-labeled training data as well as development datasets for each task
- We used 10% of the training data for validation set
- In addition, we used the unlabeled Reddit and Gab datasets, and applied silver labeling
- Finally, we used two existing public datasets:
  - EXIST
  - Call Me Sexist

Silver Labeling

- To label the provided unlabeled data, we used a weighted sum of normalized scores from two weak classifiers
    - Weights = (0.75,0.25)
- The two classifiers are:
    - RoBERTa
    - Sentence Transformer
- We also used an offensive term list, i.e. manually curated list of offensive terms.
    - Found the count of these terms in the sexist and non-sexist samples, and took the top ratios

Task A data distribution

	Train					Dev.	Test
	Gold	**Silver (Reddit)**	**Silver (Gab)**	**Call Me Sexist**	**EXIST**	**Gold**	**Gold**
Sexist	3,398	3,533	3,524	486	1,809	3,377	3,030
Non-Sexist	10,602	3,467	3,476	3,398	11,822	3,600	970

Task B data distribution

	Train		Dev.	Test
	Gold	**Silver (Reddit)**	**Gold**	**Gold**
1. Threats	310	1,000	44	89
2. Derogation	1,590	1,500	227	454
3. Animosity	1,165	2,000	167	333
4. Prejudice	333	1,000	48	94

Task C data distribution

	Train		Dev.	Test
	Gold	**Silver (Reddit)**	**Gold**	**Gold**
1.1 Threats of harm	56	500	8	16
1.2 Incitement and encouragement of harm	254	500	36	73
2.1 Descriptive attacks	717	500	102	205
2.2 Aggressive and emotive attacks	673	500	96	192
2.3 Dehumanising attacks and overt sexual objectification	200	500	29	57
3.1 Casual use of gendered slurs, profanities and insults	637	500	91	182
3.2 Immutable gender differences and gender stereotypes	417	500	60	119
3.3 Backhanded gendered compliments	64	500	9	18
3.4 Condescending explanations or unwelcome advice	47	500	7	14
4.1 Supporting mistreatment of individual women	75	500	11	21
4.2 Supporting systemic discrimination against women as a group	258	500	37	73

Results on the Development Set

Task A

Model	Non-sexist	Sexist	Macro F1
Majority Baseline	0.86	0.0	0.43
(A) BERT (L)	0.90	0.66	0.78
(B) ELECTRA (S)	0.91	0.69	0.80
(C) RoBERTa (B)	0.92	0.74	0.83
(D) RoBERTa (L)	0.92	0.70	0.81
(E) RoBERTa (L)	0.93	0.76	0.84
(F) RoBERTa (L)	0.92	0.74	0.83
(G) RoBERTa (L)	0.91	0.72	0.82
(H) RoBERTa (L)	0.93	0.75	0.84
(I) RoBERTa (L)	0.92	0.76	0.84
(J) RoBERTa (L)	0.92	0.73	0.82
(K) RoBERTa (L)	0.92	0.75	0.83
(L) RoBERTa (L)	0.93	0.76	0.85
(M) RoBERTa (L)	0.92	0.72	0.82
(N) XLM-R (L)	0.92	0.72	0.82
(O) GPT3 (Prompting)	0.85	0.61	0.73
ALL (A-O)	0.93	0.77	0.85
(E)+(F)+(H)+(I)+(L)+(O)	**0.94**	**0.80**	**0.87**

Task B

Base Model	Class 1	Class 2	Class 3	Class 4	Macro F1
Majority Baseline	0.0	0.64	0.0	0.0	0.16
(A) BERT (B)	0.66	0.71	0.58	0.56	0.63
(B) ELECTRA (L)	0.70	0.70	0.56	0.52	0.62
(C) RoBERTa (B)	0.61	0.66	0.55	0.64	0.62
(D) RoBERTa (L)	**0.75**	0.73	0.62	0.70	0.70
(E) RoBERTa (L)	0.69	0.71	0.60	0.63	0.66
(F) XLM-R (L)	0.62	0.69	0.57	0.55	0.61
(G) BERTweet (L)	0.70	0.73	0.56	0.60	0.65
(H) MiniLM (L12)	0.54	0.57	0.58	0.58	0.57
(I) GPT3 (Prompting)	0.47	0.58	0.43	0.00	0.37
ALL (A-I)	0.63	0.73	0.64	0.64	0.66
(A)+(B)+(C)+(D)+(E)	0.72	**0.75**	**0.69**	**0.72**	**0.72**

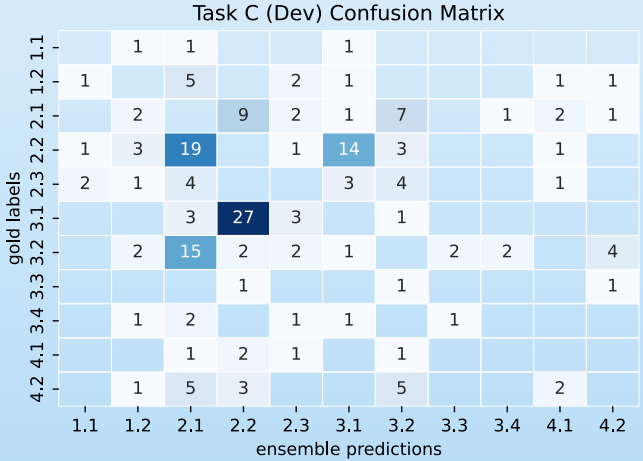
Results on the Development Set

Task C

Base Model	1.1	1.2	2.1	2.2	2.3	3.1	3.2	3.3	3.4	4.1	4.2	Macro
Majority Baseline	0.0	0.0	0.35	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.03
(A) BERT (B)	0.33	0.66	0.60	0.54	0.31	0.57	0.53	0.53	0.00	0.22	0.58	0.44
(B) ELECTRA (L)	0.44	0.59	0.61	0.51	0.45	0.55	0.45	0.32	0.00	0.27	0.54	0.52
(C) RoBERTa (B)	0.37	0.52	0.58	0.53	0.30	0.58	0.38	0.27	0.11	0.39	0.59	0.42
(D) RoBERTa (L)	0.56	**0.71**	0.61	0.50	0.43	0.61	0.51	0.53	0.22	**0.50**	**0.66**	0.53
(E) RoBERTa (L)	0.38	0.69	0.62	0.51	0.33	0.63	0.45	0.15	0.18	0.35	0.61	0.45
(F) XLM-R (L)	0.56	0.59	0.56	0.47	0.40	0.55	0.49	0.40	0.00	0.30	0.53	0.44
(G) BERTweet (L)	0.50	0.68	0.61	0.53	0.38	0.60	0.46	0.15	**0.22**	0.38	0.61	0.47
(H) MiniLM (L12)	0.56	0.46	0.45	0.28	0.24	0.58	0.45	0.25	0.13	0.44	0.51	0.40
(I) ALBERT v2 (XL)	0.50	0.50	0.47	0.43	0.31	0.49	0.44	0.40	0.00	0.36	0.43	0.39
ALL (A-I)	0.37	0.68	0.65	**0.57**	0.37	0.64	0.53	0.38	0.00	**0.50**	0.60	0.48
(B)+(D)+(E)+(F)	**0.59**	0.69	**0.66**	0.56	**0.51**	**0.67**	**0.54**	**0.67**	0.18	**0.50**	0.65	**0.56**

Error Analysis

Since the Task C categories are the most fine- grained, we investigated for which Task C categories the model under-performed, by comparing the ensemble predictions on the development set with the task C development set gold labels.



Error Analysis

The majority of the development set errors stemmed from the failure of the model to distinguish between:

- Labels 2.2 (Aggressive and emotive attacks) and 3.1 (Casual use of gendered slurs, profanities and insults)
- Labels 2.1 (Descriptive attacks) and 2.2 (Aggressive and emotive attacks)
- Labels 2.1 (Descriptive attacks) and 3.2 (Immutable gender differences and gender stereotypes)

Gold	Pred	Text
2.1	2.2	*Treat a whore like a lady, and treat a lady like a whore.*
2.2	2.1	*How do block people from following me? A bunch of nasty women are following me on gab.i'm disgusted!!*
2.2	3.1	*That,Äôs is more likely your meth whore mom,Äôs job. Go fuck yourself, Nazi fag.*
3.1	2.2	*I'd rather have a loyal friend than a dumb thot*
3.2	2.1	*Do you want answers or not? You asked trp and we answered. Just because you don't like the answer doesn't make it false - that's a woman's line of think-ing.*

Conclusion

Detection of online sexism is important, but being able to explain the type of sexism is even more important.

Through our experiments, we found that:

- Utilizing a combination of data augmentation methods is helpful
  - Publicly available datasets
  - Silver labeling
- Using ensembling of different transformer-based models gives better performance than using a single model
- In-context learning using GPT3 was helpful for Task A, but no Task B and C

Thank you for your time

troosta@amazon.com