Natural Language Processing and Information Extraction, 2024WS

17.01.2025

Sexism Detection Project

presented by the

Label Legends

Peter Reti (12432931)
Lukas Mahler (11908553)

Noemi Gazdik (12432929)
Jacopo Raffaelli (12329537)

Agenda

01 Topic and Task Description

02 First Milestone

03 Second Milestone

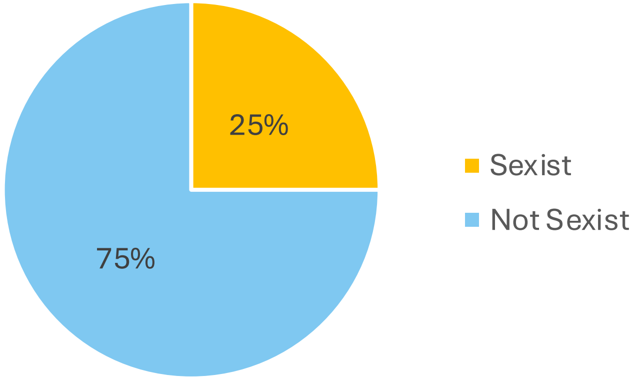
04 Final Milestone

01

Topic and Task Description

Introduction

- About out Topic
  - Topic 4 – Detection of Online Sexism
  - The goal of our task – Classification model to detect sexism on social media posts
  - Basis – "SemEval-2023 Task 10: Explainable Detection of Online Sexism" (Kirk et al., 2023)
- About our Dataset
  - Reddit comments, Gab posts
  - 60k rows (split already determined)
    - 40k rows in the training dataset
    - 20k rows in the test dataset
    - 75% non-sexist
  - Each text appears 3 times, labeled by 3 different annotators (out of 19)
  - The annotators are highly-trained and all self-identify as women



Task Description

- Set-up & Preprocessing
- Baseline Models

  Q Which model is better suited to the task, and why?

- Error Analysis

  Q Do you find any lexical or linguistic patterns in general that always correspond with sexist content?

- Improvements ideas
- Final Model

02

First Milestone

Setup & Preprocessing

- uv for dependency management
- Central preprocessing
  - Each model recieves preprocessed data
- Convert labels to integer
- Dropped multiclass labels
- Clean masked our data
  - e.g. "[URL]", "[CLICK HERE]","[USER]"

Lemmatization

- stanza library
- Persisted in CoNNL-U format
- Used by
  - Regex Baseline
  - XGBoost
  - Ensemble improvements for DeBERTa

03

Second Milestone

About our models

Baseline models

Non-Deep Learning Models

- Most Frequent
- Regex
- XGBoost

Deep Learning Models

- DistillBERT
- DeBERTa-v3-base

Metrics used for evaluation

- Macro AVG F1
- Precision
- Recall
- Training Time
- Test Time

Baseline results

Metrics	Most Frequent Baseline	Regex Baseline	XGBoost Baseline	DistilBERT Baseline	DeBERTa-v3-base baseline	XGBoost hyperopt. (GPU)	DistilBERT 50 epochs (GPU)
Precision	0.0000	0.3414	0.7639	0.7237	0.7235	0.7254	0.6872
Recall	0.0000	0.6087	0.3884	0.5583	0.6738	0.4520	0.6453
F-Score	0.0000	0.4375	0.5150	0.6303	0.6978	0.5570	0.6656
Accuracy	0.7404	0.5937	0.8101	0.8300	0.8485	0.8133	0.8317
Train time	0.00 s	3.72 s	2.88 s	1538.93 s	4893.40 s	1800 s	10902 s
Test time	0.00 s	6.84 s	0.13 s	52.21 s	132.74 s	0.13 s	42.84 s

Best model: DeBERTa

Error Analysis

Goal

- thorough qualitative analysis
- identify key challanges

Focus

There are much more meaningful steps we can take if we want to catch **false negatives.**

Steps

1. Dataframe: Text, True label, Predictions (5)

2. Assigning Error levels to the Texts = how many models predicted it **wrong**

   0 – best, 5 - worst

Error level	n
0	4659
5	474
4-5	1269

Result of our Error Analysis

Error levels 4 -5

We chose this category because we think this way we can see what sentences/words made it hard for the models to predict right.

→ Findings:

- Personal pronouns disappear during our tokenization.
- Hard to detecting sarcasm or stereotype connected to the female sex.
- The models struggle with texts containing slang.
- In cases the annotators did not agree – the model gets this at least once wrong

Solution ideas

1. Merge all instances of a text together, edit label based on majority
2. New label based on these functions

   Is the text gender (in our case female) specific?

   Does the text contain bad words?

   Does the text have negative annotation?
3. Collecting derogatory words against women from error level 5, increasing the weights of these tokens in the model input
4. Introduce weights based on annotator sensitivity
5. New tokenization include e.g. he, she, shortened denial words (don't), slang words
6. Add phrases that were hard to predict as sexist in the train dataset

04 Final Milestone

Our approach

Our solutions should offer flexible improvement not just case specific

+

Our solution should be scalable

+

Our solution should focus on Recall

Final solution idea

~~Preprocessing step~~

- ~~Merge all instances of a text together and edit label based on majority votes of it being sexist or not~~

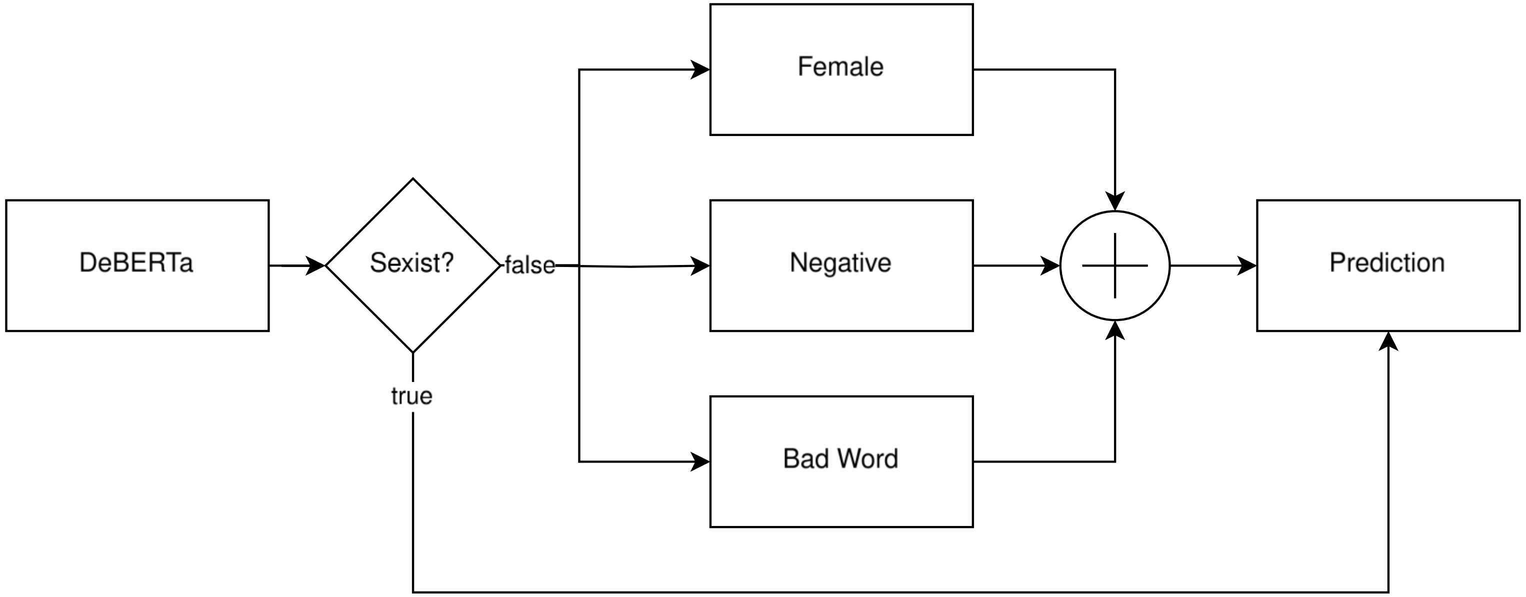
Post Processing step

- Function deciding whether a text is gender (in our case female) specific
- Function deciding whether the text has negative annotation or not
- Function deciding whether the text contains bad words or not

Model	Precision	Recall	F1 Score	Accuracy	TN	FP	FN	TP
DeBERTa Base	0.7213	0.6530	0.6854	0.8444	8099	786	1081	2034
DeBERTa Majority	0.7483	0.5859	0.6572	0.8413	8271	614	1290	1825

Ensemble model

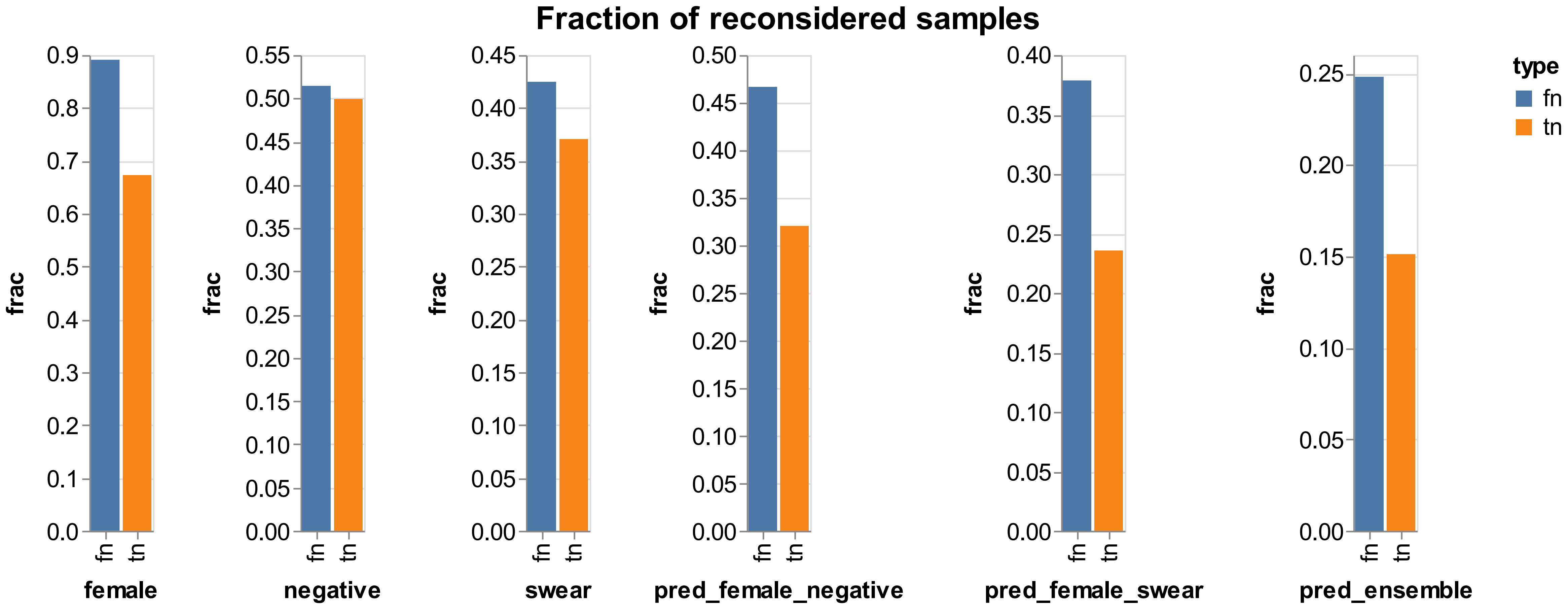
- Based on DeBERTa
- Post-Process negative predictions using
  - rule-based system
  - Sentiment analysis
  - Database of "bad" words
- Goal: Increase Recall



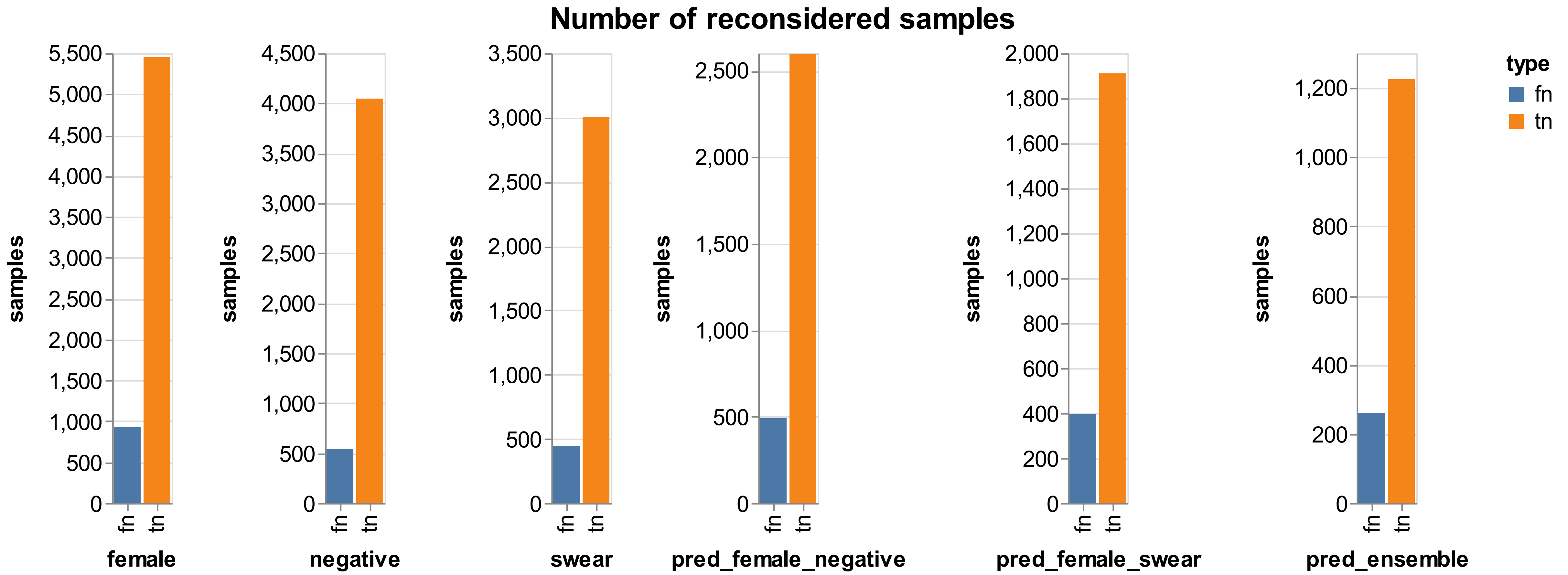
Ensemble model

- If DeBERTa predicts negative:
  - Detect if a "female specific" word appears in the text
    - Rule-based system
    - e.g. woman, girl, wife
    - Created by us based on validation set
  - Detect if the text has negative connotation
    - Sentiment analysis
    - VaderSentiment
  - Detect if a "bad" word appears in the text
    - Slurs, derogatory terms
    - Used following dataset: https://www.kaggle.com/datasets/tushifire/ldnoobw

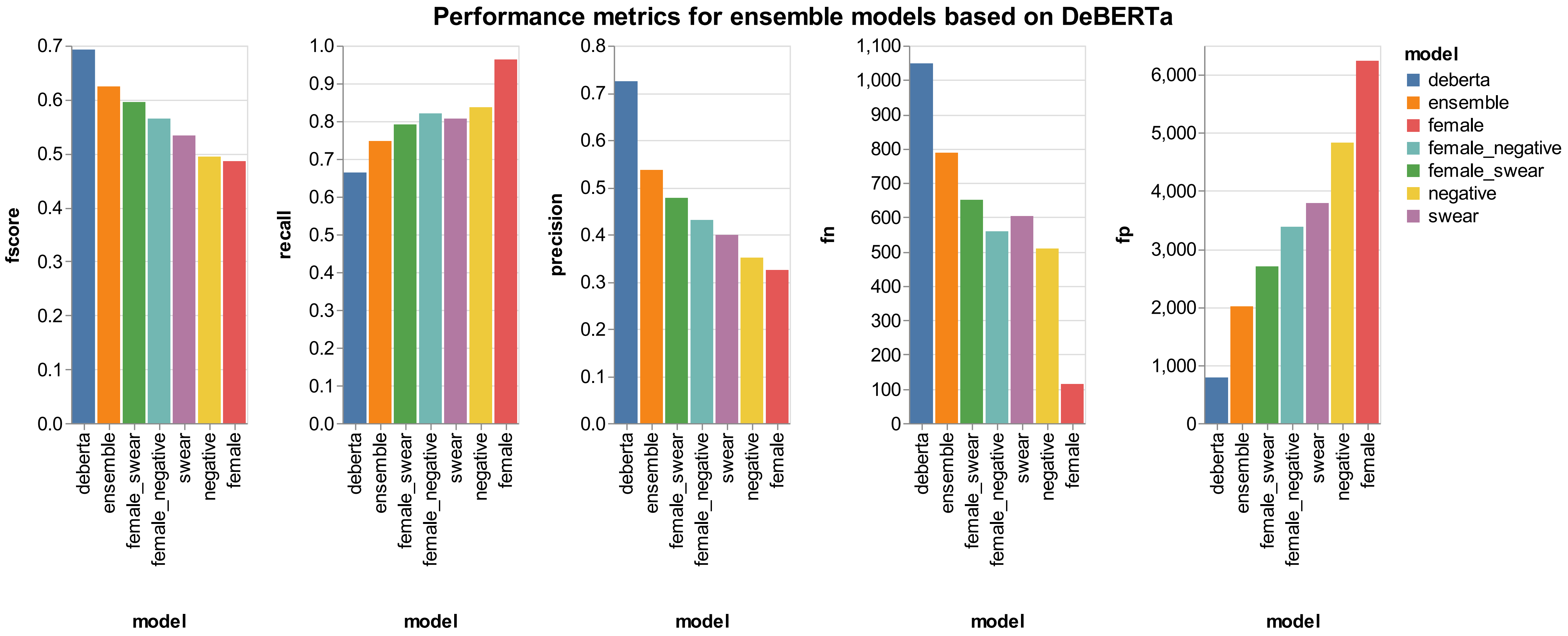
Ensemble model



Ensemble model



Results


Performance metrics for ensemble models based on DeBERTa
fscore
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
recall
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
precision
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
fn
0
100
200
300
400
500
600
700
800
900
1,000
1,100
fp
0
1,000
2,000
3,000
4,000
5,000
6,000
deberta
ensemble
female_swear
female_negative
swear
negative
female
model
model
deberta
ensemble
female
female_negative
female_swear
negative
swear

Final Solution

See if the code is clean

Management summary

Documentation

Test the model on another dataset ?

Thank you for your attention!

Any Questions or Suggestions?

References

[1] Hannah Kirk et al. “SemEval-2023 Task 10: Explainable Detection of Online Sexism”. In: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval 2023). Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 2193–2210. doi: 10.18653/v1/2023.semeval-1.305. url: https://aclanthology.org/2023.semeval-1.305.