TECHNISCHE
UNIVERSITÄT
WIEN

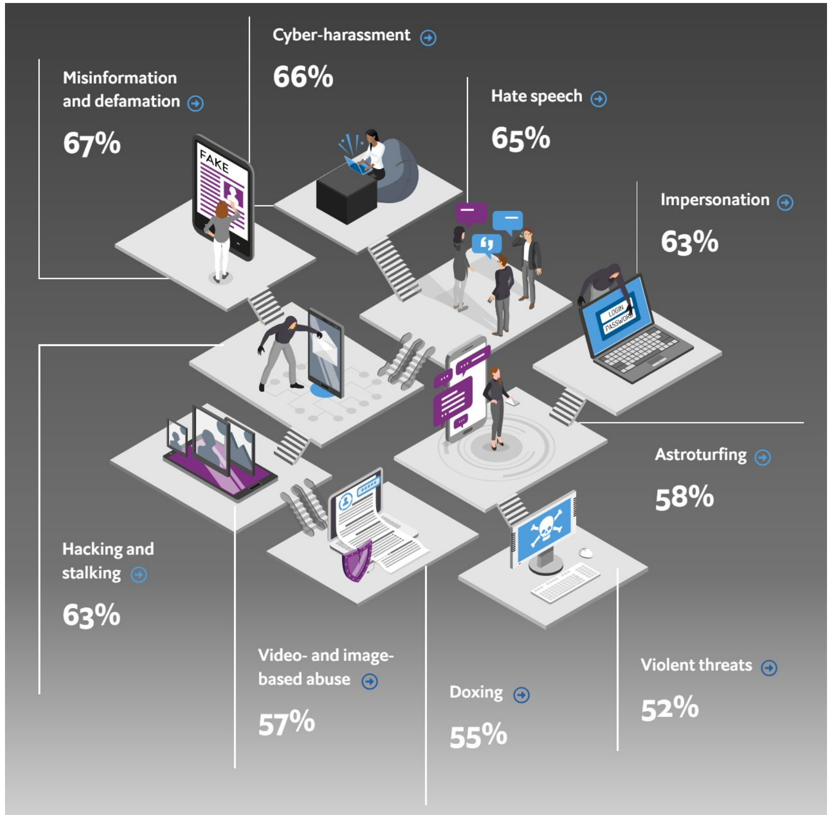
Online Sexism Detection

194.093 Natural Language Processing and Information Extraction

Draganic Vanja	| 12308447
Milicic Sofija	| 12239374
Forman Henry	| 11803113
Pfennigbauer Johannes	| 11902046

Background

- 85% of women reported witnessing online violence against other women
- 38% of women reported personal experience
- Women are 27 times more likely to face harassment online than men
- 62% experience a sense of helplessness as little is done to combat the issue



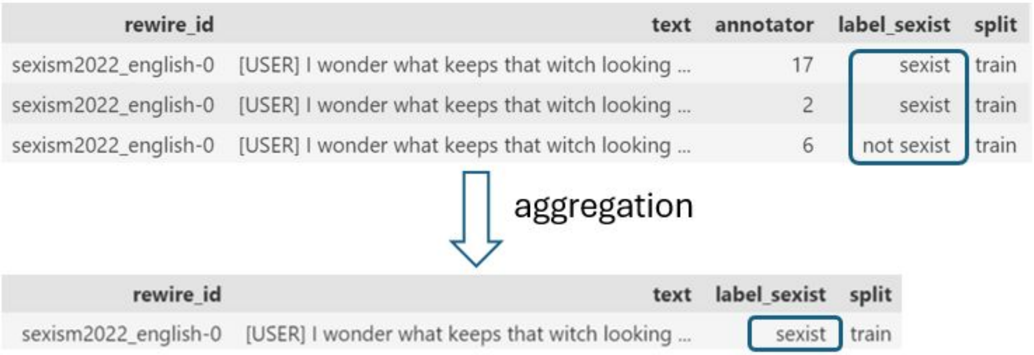
Data refers to overall prevalence rates by threat tactic, in %.
Source: https://onlineviolencewomen.eiu.com/

Research Question

- How can (advanced) natural language processing models, combined with effective explainability techniques, be leveraged to improve the detection of online sexism?

(The Economist Intelligence Unit, 2021)

Dataset



rewire_id	text	annotator	label_sexist	split
sexism2022_english-0	[USER] I wonder what keeps that witch looking ...	17	sexist	train
sexism2022_english-0	[USER] I wonder what keeps that witch looking ...	2	sexist	train
sexism2022_english-0	[USER] I wonder what keeps that witch looking ...	6	not sexist	train

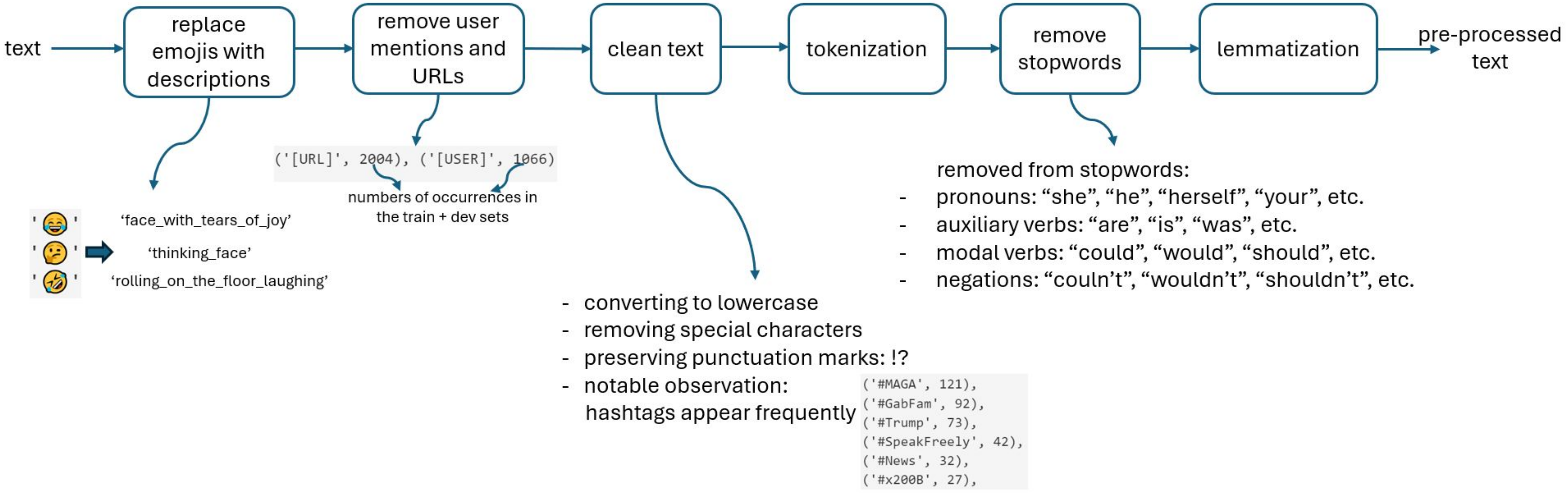
aggregation

rewire_id	text	label_sexist	split
sexism2022_english-0	[USER] I wonder what keeps that witch looking ...	sexist	train

- 2 versions of the dataset
  1. individual labels
     - each comment annotated by 3 different annotators
     - total: 60,000 entries (20,000 comments x 3 annotators)
     - full agreement (3/3): achieved in most cases
     - partial agreement (2/3): occurred in 4,444 cases
  2. aggregated labels
     - partial (2/3) agreements resolved by dataset creators
     - total: 20,000 entries (single label per comment)
- comments collected from 2 social media platforms: Gab and Reddit

(Kirk et al., 2023)

Data Preprocessing



text → replace emojis with descriptions → remove user mentions and URLs → clean text → tokenization → remove stopwords → lemmatization → pre-processed text

'😂' 'face_with_tears_of_joy'
'🤔' ➡ 'thinking_face'
'🤣' 'rolling_on_the_floor_laughing'

```
('[URL]', 2004), ('[USER]', 1066)
```

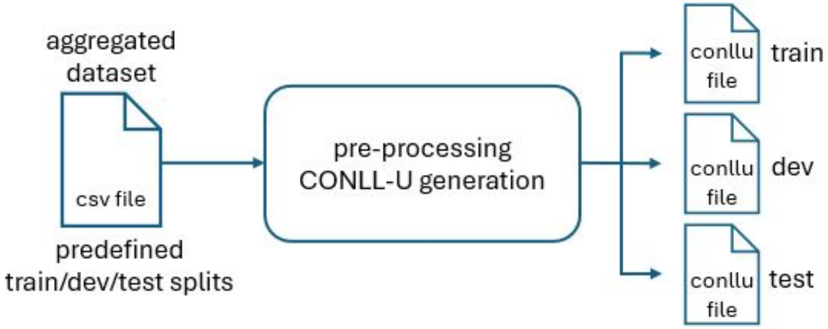
numbers of occurrences in the train + dev sets

- converting to lowercase
- removing special characters
- preserving punctuation marks: !?
- notable observation:
  hashtags appear frequently

```
('#MAGA', 121),
('#GabFam', 92),
('#Trump', 73),
('#SpeakFreely', 42),
('#News', 32),
('#x200B', 27),
```

removed from stopwords:

- pronouns: “she”, “he”, “herself”, “your”, etc.
- auxiliary verbs: “are”, “is”, “was”, etc.
- modal verbs: “could”, “would”, “should”, etc.
- negations: “couln’t”, “wouldn’t”, “shouldn’t”, etc.



aggregated dataset

csv file

predefined train/dev/test splits

pre-processing CONLL-U generation

conllu file train

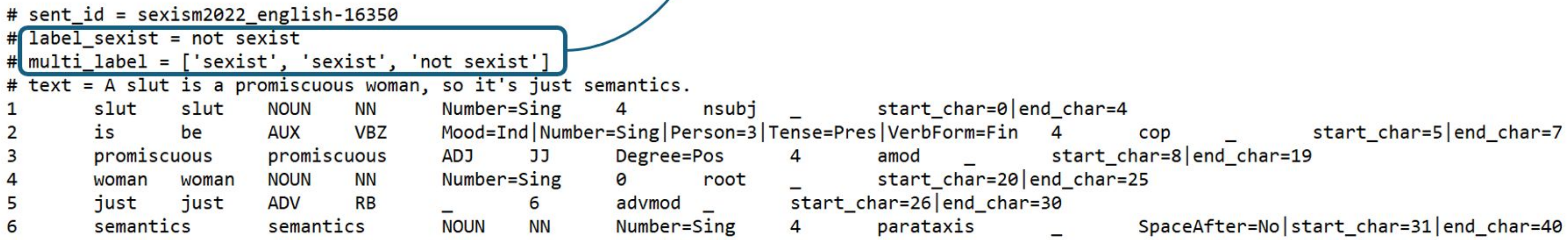
conllu file dev

conllu file test

Label Aggregation

different aggregation types:

1. original → “not sexist”
2. majority voting → “sexist”
3. at least one sexist → “sexist”



```
# sent_id = sexism2022_english-16350
# label_sexist = not sexist
# multi_label = ['sexist', 'sexist', 'not sexist']
# text = A slut is a promiscuous woman, so it's just semantics.
1       slut    slut    NOUN    NN      Number=Sing     4       nsubj   _       start_char=0|end_char=4
2       is      be      AUX     VBZ     Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   4       cop     _       start_char=5|end_char=7
3       promiscuous     promiscuous     ADJ     JJ      Degree=Pos      4       amod    _       start_char=8|end_char=19
4       woman   woman   NOUN    NN      Number=Sing     0       root    _       start_char=20|end_char=25
5       just    just    ADV     RB      _       6       advmod  _       start_char=26|end_char=30
6       semantics       semantics       NOUN    NN      Number=Sing     4       parataxis       _       SpaceAfter=No|start_char=31|end_char=40
```

Baseline Models

Model	**Balanced Accuracy**	**Recall**
Majority class	50,0%	00,0%
Naive Bayes	69,1%	63,8%
Logistic Regression	63,7%	40,0%
XGBoost (BOW)	77,8%	58,2%
XGBoost (TF-IDF)	78,3%	55,3%
LSTM	73,2%	64,2%

Results of different baseline models trained on the balanced training set, evaluated on the validation dataset.

Results for different types of aggregation

Model	Aggregation Type	Accuracy	Balanced Accuracy	Precision	Recall
LSTM	At least one sexist	72,2%	62,3%	80,9%	28,1%
	Majority voting	83,2%	68,4%	79,3%	40,1%
	Original	81,5%	65,3%	77,4%	33,7%
Naive Bayes	At least one sexist	72,4%	69,0%	61,9%	57,3%
	Majority voting	81,6%	72,0%	63,8%	53,7%
	Original	81,0%	71,3%	63,0%	52,5%

Results of selected baseline models with different aggregation techniques evaluated on the validation dataset.

BERT-based Models

- **contextual representation** → better performance than baseline models
- models experimented with:
  1. RoBERTa
     - dynamic masking, no NSP, only MLM, trained on 10x more data
     - HF model fine-tuned for sentiment analysis on Twitter → comparable results to HateBERT
  2. DeBERTa
     - disentangled attention (separate vectors for content & position)
  3. **HateBERT**
     - fine-tuned on a large dataset of banned Reddit communities (RAL-E dataset; offensive, abusive or hateful speech)
     - strength: specialized for hate speech, weakness: potential for bias (false positives)
  4. DistilBERT
     - smaller, faster and more efficient version of the BERT, uses knowledge distillation
- fine-tuning implementation details: **PyTorch training loop** (epoch by epoch, batch by batch…), **Hugging Face's Trainer API** (easier to use, but with longer training time)

BERT-based Models Results

TECHNISCHE
UNIVERSITÄT
WIEN

Model	Accuracy	Balanced Accuracy	Precision	Recall
RoBERTa	84,9%	80,1%	68,4%	70,8%
DeBERTa	75,7%	50,0%	00,0%	00,0%
HateBERT	86,2%	78,6%	75,8%	63,8%
DistilBERT	83,6%	74,7%	69,9%	57,4%

Results of different BERT-based trained on **original training set** models evaluated on the validation dataset

Model	Aggreg. Type	Accuracy	Balanced Accuracy	Precision	Recall
Hate BERT	At least one sexist	79,1%	77,2%	70,6%	70,5%
	Majority voting	85,1%	77,6%	71,7%	62,8%
	Original	86,1%	77,7%	76,8%	61,3%

Results of HateBERT with different aggregation techniques, trained on the **original training set**, evaluated on the validation set

Model	Aggreg. Type	Accuracy	Balanced Accuracy	Precision	Recall
Hate BERT	At least one sexist	78,0%	77,6%	66,6%	76,3%
	Majority voting	84,5%	77,0%	69,9%	62,4%
	Original	81,4%	79,9%	59,0%	77,0%

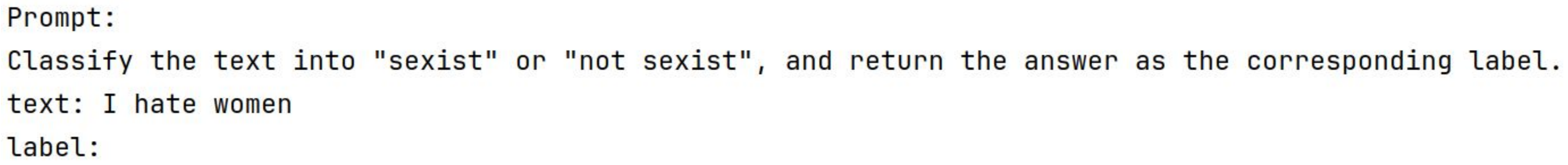
Results of HateBERT with different aggregation techniques, trained on the **balanced training set**, evaluated on the validation set

Llama 3.2 3B Instruct

- LLM developed by Meta
- Text to text model
- Instruction-tuned
- 3.24 billion parameters

Llama Setup

- Raw sentence as input
- Passing sentence as shown below
- Low temperature
- Generated 2 additional tokens
- Additional class: “none”
- Fine-tuned on provided aggregation with train and balanced data set



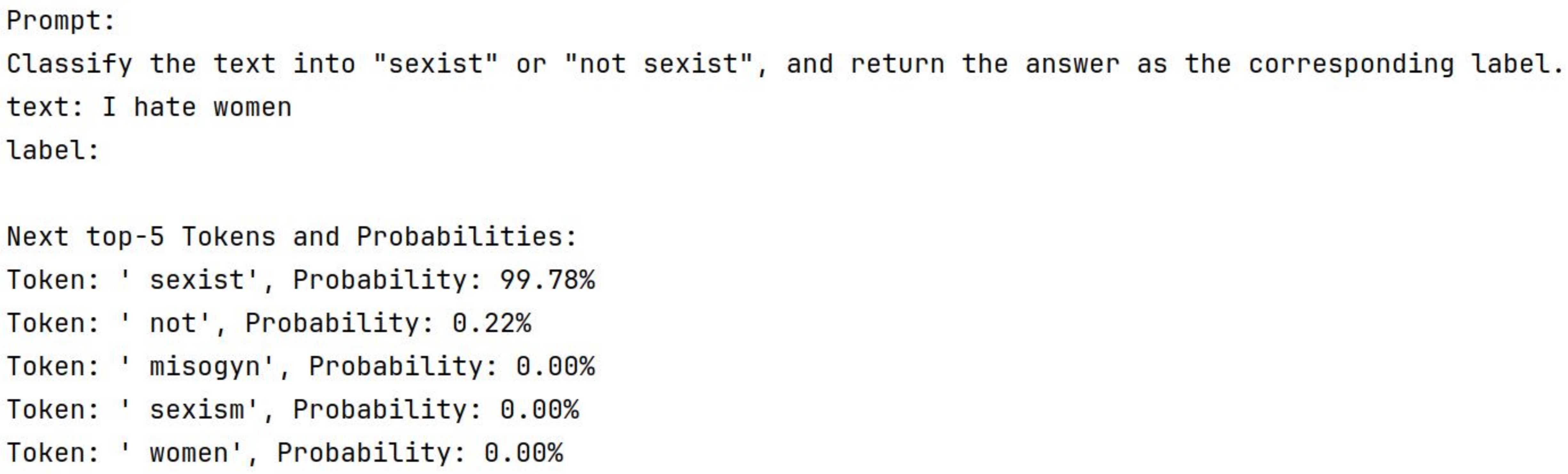
```
Prompt:
Classify the text into "sexist" or "not sexist", and return the answer as the corresponding label.
text: I hate women
label:
```

Llama Results

Model	**Aggregation Type**	**Fine-tuned on**	**Accuracy**	**Balanced Accuracy**	**Precision**	**Recall**
LSTM	Original	-	81,5%	65,3%	77,4%	33,7%
Naive Bayes	Original	-	81,0%	71,3%	63,0%	52,5%
HateBERT	Original	train	86,2%	78,6%	75,8%	63,8%
Llama 3.2 3B	Original	-	58,5%	66,6%	35,1%	82,5%
Llama 3.2 3B	Original	train data	**88,3%**	**86,0%**	73,5%	**81,5%**
Llama 3.2 3B	Original	balanced data	88,4%	80,8%	**83,1%**	65,%

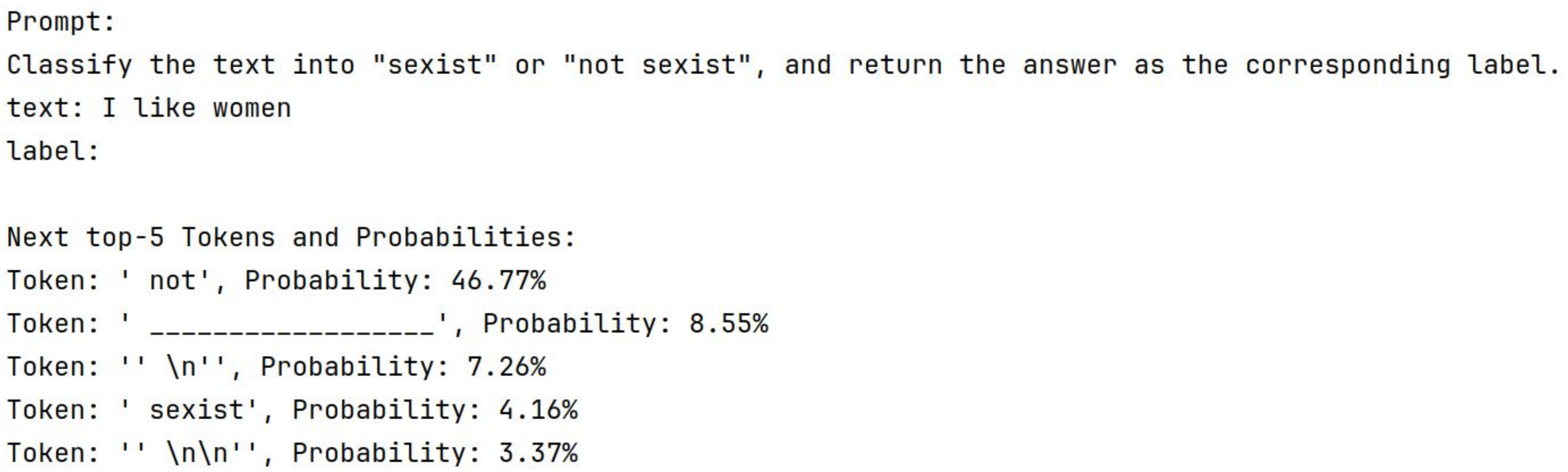
Results of selected models evaluated on the original validation dataset. Precision and recall are calculated for the label “sexist”

Llama Token probabilities



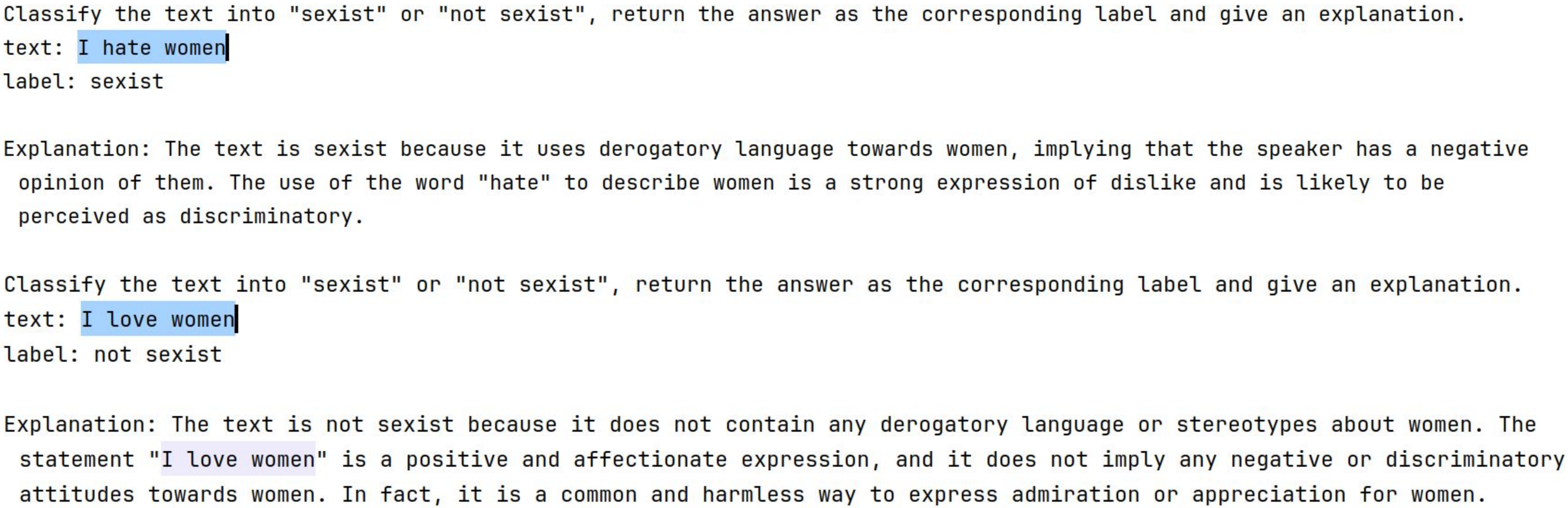
```
Prompt:
Classify the text into "sexist" or "not sexist", and return the answer as the corresponding label.
text: I hate women
label:

Next top-5 Tokens and Probabilities:
Token: ' sexist', Probability: 99.78%
Token: ' not', Probability: 0.22%
Token: ' misogyn', Probability: 0.00%
Token: ' sexism', Probability: 0.00%
Token: ' women', Probability: 0.00%
```

Llama Token probabilities



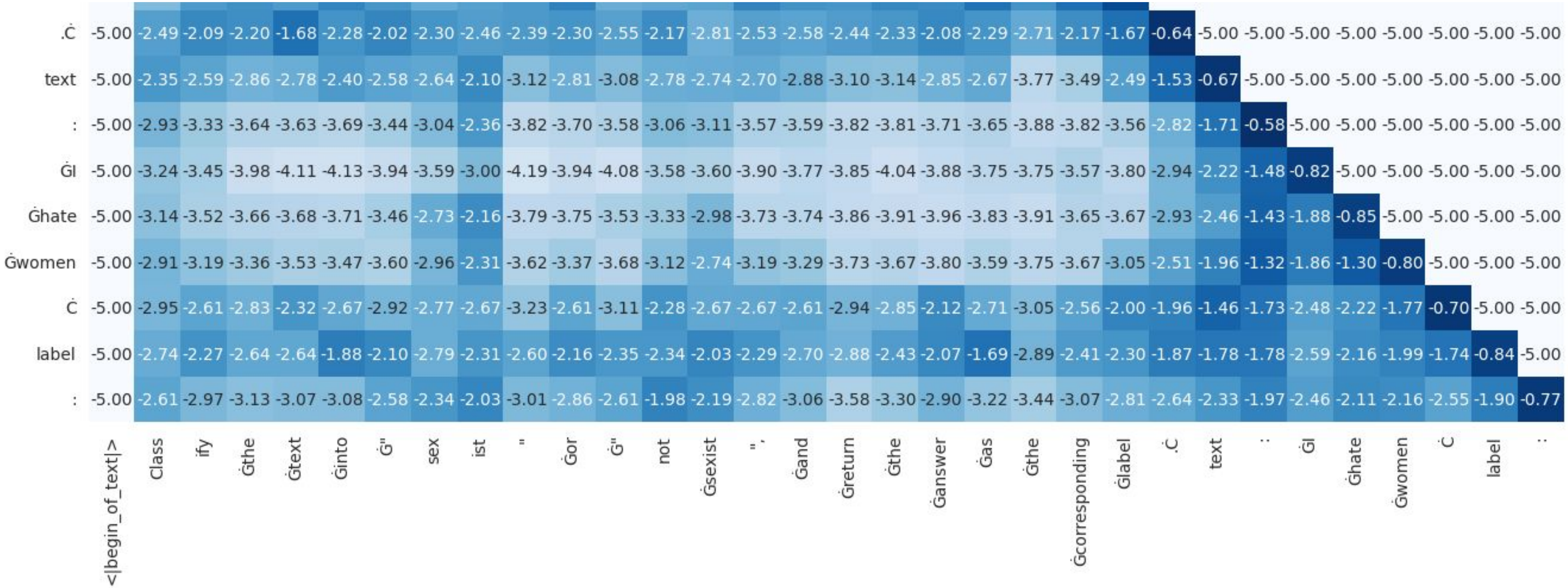
```
Prompt:
Classify the text into "sexist" or "not sexist", and return the answer as the corresponding label.
text: I like women
label:

Next top-5 Tokens and Probabilities:
Token: ' not', Probability: 46.77%
Token: ' _________________', Probability: 8.55%
Token: '' \n'', Probability: 7.26%
Token: ' sexist', Probability: 4.16%
Token: '' \n\n'', Probability: 3.37%
```

Llama Classification Explanation



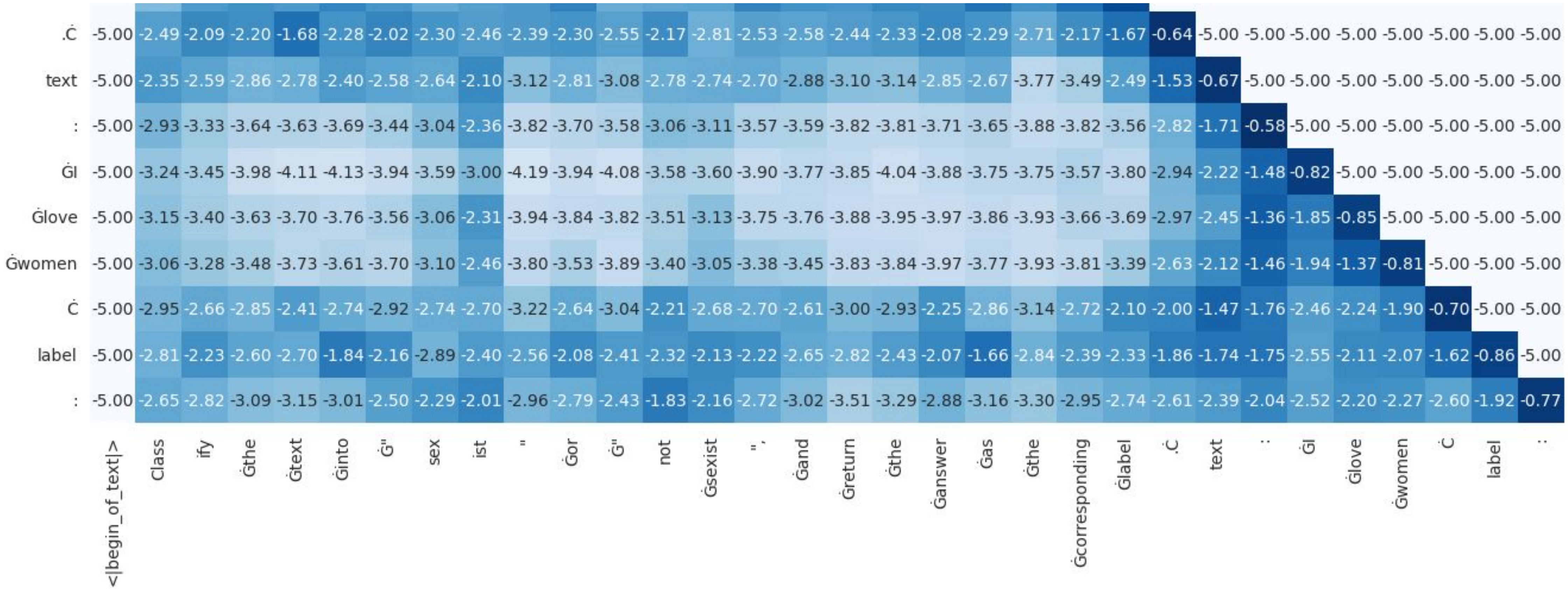
```
Classify the text into "sexist" or "not sexist", return the answer as the corresponding label and give an explanation.
text: I hate women
label: sexist

Explanation: The text is sexist because it uses derogatory language towards women, implying that the speaker has a negative
 opinion of them. The use of the word "hate" to describe women is a strong expression of dislike and is likely to be
 perceived as discriminatory.

Classify the text into "sexist" or "not sexist", return the answer as the corresponding label and give an explanation.
text: I love women
label: not sexist

Explanation: The text is not sexist because it does not contain any derogatory language or stereotypes about women. The
 statement "I love women" is a positive and affectionate expression, and it does not imply any negative or discriminatory
 attitudes towards women. In fact, it is a common and harmless way to express admiration or appreciation for women.
```

Llama Attention Heatmap



	<|begin_of_text|>	Class	ify	Ġthe	Ġtext	Ġinto	Ġ"	sex	ist	"	Ġor	Ġ"	not	Ġsexist	",	Ġand	Ġreturn	Ġthe	Ġanswer	Ġas	Ġthe	Ġcorresponding	Ġlabel	.Ċ	text	:	ĠI	Ġhate	Ġwomen	Ċ	label	:
.Ċ	-5.00	-2.49	-2.09	-2.20	-1.68	-2.28	-2.02	-2.30	-2.46	-2.39	-2.30	-2.55	-2.17	-2.81	-2.53	-2.58	-2.44	-2.33	-2.08	-2.29	-2.71	-2.17	-1.67	-0.64	-5.00	-5.00	-5.00	-5.00	-5.00	-5.00	-5.00	-5.00
text	-5.00	-2.35	-2.59	-2.86	-2.78	-2.40	-2.58	-2.64	-2.10	-3.12	-2.81	-3.08	-2.78	-2.74	-2.70	-2.88	-3.10	-3.14	-2.85	-2.67	-3.77	-3.49	-2.49	-1.53	-0.67	-5.00	-5.00	-5.00	-5.00	-5.00	-5.00	-5.00
:	-5.00	-2.93	-3.33	-3.64	-3.63	-3.69	-3.44	-3.04	-2.36	-3.82	-3.70	-3.58	-3.06	-3.11	-3.57	-3.59	-3.82	-3.81	-3.71	-3.65	-3.88	-3.82	-3.56	-2.82	-1.71	-0.58	-5.00	-5.00	-5.00	-5.00	-5.00	-5.00
ĠI	-5.00	-3.24	-3.45	-3.98	-4.11	-4.13	-3.94	-3.59	-3.00	-4.19	-3.94	-4.08	-3.58	-3.60	-3.90	-3.77	-3.85	-4.04	-3.88	-3.75	-3.75	-3.57	-3.80	-2.94	-2.22	-1.48	-0.82	-5.00	-5.00	-5.00	-5.00	-5.00
Ġhate	-5.00	-3.14	-3.52	-3.66	-3.68	-3.71	-3.46	-2.73	-2.16	-3.79	-3.75	-3.53	-3.33	-2.98	-3.73	-3.74	-3.86	-3.91	-3.96	-3.83	-3.91	-3.65	-3.67	-2.93	-2.46	-1.43	-1.88	-0.85	-5.00	-5.00	-5.00	-5.00
Ġwomen	-5.00	-2.91	-3.19	-3.36	-3.53	-3.47	-3.60	-2.96	-2.31	-3.62	-3.37	-3.68	-3.12	-2.74	-3.19	-3.29	-3.73	-3.67	-3.80	-3.59	-3.75	-3.67	-3.05	-2.51	-1.96	-1.32	-1.86	-1.30	-0.80	-5.00	-5.00	-5.00
Ċ	-5.00	-2.95	-2.61	-2.83	-2.32	-2.67	-2.92	-2.77	-2.67	-3.23	-2.61	-3.11	-2.28	-2.67	-2.67	-2.61	-2.94	-2.85	-2.12	-2.71	-3.05	-2.56	-2.00	-1.96	-1.46	-1.73	-2.48	-2.22	-1.77	-0.70	-5.00	-5.00
label	-5.00	-2.74	-2.27	-2.64	-2.64	-1.88	-2.10	-2.79	-2.31	-2.60	-2.16	-2.35	-2.34	-2.03	-2.29	-2.70	-2.88	-2.43	-2.07	-1.69	-2.89	-2.41	-2.30	-1.87	-1.78	-1.78	-2.59	-2.16	-1.99	-1.74	-0.84	-5.00
:	-5.00	-2.61	-2.97	-3.13	-3.07	-3.08	-2.58	-2.34	-2.03	-3.01	-2.86	-2.61	-1.98	-2.19	-2.82	-3.06	-3.58	-3.30	-2.90	-3.22	-3.44	-3.07	-2.81	-2.64	-2.33	-1.97	-2.46	-2.11	-2.16	-2.55	-1.90	-0.77

Llama Attention Heatmap



	<|begin_of_text|>	Class	ify	Ġthe	Ġtext	Ġinto	Ġ"	sex	ist	"	Ġor	Ġ"	not	Ġsexist	",	Ġand	Ġreturn	Ġthe	Ġanswer	Ġas	Ġthe	Ġcorresponding	Ġlabel	.Ċ	text	:	ĠI	Ġlove	Ġwomen	Ċ	label	:
.Ċ	-5.00	-2.49	-2.09	-2.20	-1.68	-2.28	-2.02	-2.30	-2.46	-2.39	-2.30	-2.55	-2.17	-2.81	-2.53	-2.58	-2.44	-2.33	-2.08	-2.29	-2.71	-2.17	-1.67	-0.64	-5.00	-5.00	-5.00	-5.00	-5.00	-5.00	-5.00	-5.00
text	-5.00	-2.35	-2.59	-2.86	-2.78	-2.40	-2.58	-2.64	-2.10	-3.12	-2.81	-3.08	-2.78	-2.74	-2.70	-2.88	-3.10	-3.14	-2.85	-2.67	-3.77	-3.49	-2.49	-1.53	-0.67	-5.00	-5.00	-5.00	-5.00	-5.00	-5.00	-5.00
:	-5.00	-2.93	-3.33	-3.64	-3.63	-3.69	-3.44	-3.04	-2.36	-3.82	-3.70	-3.58	-3.06	-3.11	-3.57	-3.59	-3.82	-3.81	-3.71	-3.65	-3.88	-3.82	-3.56	-2.82	-1.71	-0.58	-5.00	-5.00	-5.00	-5.00	-5.00	-5.00
ĠI	-5.00	-3.24	-3.45	-3.98	-4.11	-4.13	-3.94	-3.59	-3.00	-4.19	-3.94	-4.08	-3.58	-3.60	-3.90	-3.77	-3.85	-4.04	-3.88	-3.75	-3.75	-3.57	-3.80	-2.94	-2.22	-1.48	-0.82	-5.00	-5.00	-5.00	-5.00	-5.00
Ġlove	-5.00	-3.15	-3.40	-3.63	-3.70	-3.76	-3.56	-3.06	-2.31	-3.94	-3.84	-3.82	-3.51	-3.13	-3.75	-3.76	-3.88	-3.95	-3.97	-3.86	-3.93	-3.66	-3.69	-2.97	-2.45	-1.36	-1.85	-0.85	-5.00	-5.00	-5.00	-5.00
Ġwomen	-5.00	-3.06	-3.28	-3.48	-3.73	-3.61	-3.70	-3.10	-2.46	-3.80	-3.53	-3.89	-3.40	-3.05	-3.38	-3.45	-3.83	-3.84	-3.97	-3.77	-3.93	-3.81	-3.39	-2.63	-2.12	-1.46	-1.94	-1.37	-0.81	-5.00	-5.00	-5.00
Ċ	-5.00	-2.95	-2.66	-2.85	-2.41	-2.74	-2.92	-2.74	-2.70	-3.22	-2.64	-3.04	-2.21	-2.68	-2.70	-2.61	-3.00	-2.93	-2.25	-2.86	-3.14	-2.72	-2.10	-2.00	-1.47	-1.76	-2.46	-2.24	-1.90	-0.70	-5.00	-5.00
label	-5.00	-2.81	-2.23	-2.60	-2.70	-1.84	-2.16	-2.89	-2.40	-2.56	-2.08	-2.41	-2.32	-2.13	-2.22	-2.65	-2.82	-2.43	-2.07	-1.66	-2.84	-2.39	-2.33	-1.86	-1.74	-1.75	-2.55	-2.11	-2.07	-1.62	-0.86	-5.00
:	-5.00	-2.65	-2.82	-3.09	-3.15	-3.01	-2.50	-2.29	-2.01	-2.96	-2.79	-2.43	-1.83	-2.16	-2.72	-3.02	-3.51	-3.29	-2.88	-3.16	-3.30	-2.95	-2.74	-2.61	-2.39	-2.04	-2.52	-2.20	-2.27	-2.60	-1.92	-0.77

Error Analysis-Baseline Models

False Positives (FP)

Characteristics:

- Sentences incorrectly classified as sexist, possibly due to:
    - The presence of controversial or vulgar language.
    - Discussions of gender dynamics in neutral or balanced ways.
    - General references to feminists, MGTOW, or gendered terms (e.g., “Trouble with MGTOW is you take isolated incidents and judge all females like this”).

Patterns:

- Sentences discussing gender-related topics neutrally.
- Vulgar language unrelated to sexism (“Check the constitution, mother fucker”).
- Balanced critique or reflections on societal issues involving gender.

Examples:

- “You can’t get women because you have a shit personality. Maybe you are just creepy and it turns women away.”
- “Women should be the first test group. No. They don’t think that women should be the ‘first test group.’”

False Negatives (FN)

Characteristics:

- Sentences that contain sexism but were misclassified as not sexist.
- Often include:
    - Subtle or euphemistic sexism (e.g., “Women seemingly have endless amounts of time to waste”).
    - Stereotypical gender roles (e.g., “Women talk for the sake of talking”).
    - Sarcasm or coded language (e.g., “Chad is also catching regret rape claims later”).

Patterns:

- Subtle misogyny masked by indirect or sarcastic language.
- Stereotypical comments that lack explicit slurs but imply sexism.
- Euphemistic or metaphorical references to sexism.

Examples:

- “This shows that there ain’t a limit to stupidity for females.”
- “The boys should be killed. OK, now that we have that out of the way, why is a girl meeting two boys in a ditch to buy a knife?”

LSTM predictions

I think women belong in the kitchen -> sexist

I don’t think women belong in the kitchen -> not sexist

I think forks belong in the kitchen -> not sexist

Women should have the right to vote -> sexist

Women shouldn’t have the right to vote -> sexist

HateBERT predictions

I think women belong in the kitchen -> sexist

I don’t think women belong in the kitchen -> sexist

I think forks belong in the kitchen -> not sexist

Women should have the right to vote -> sexist

Women shouldn’t have the right to vote -> sexist

LLAMA predictions

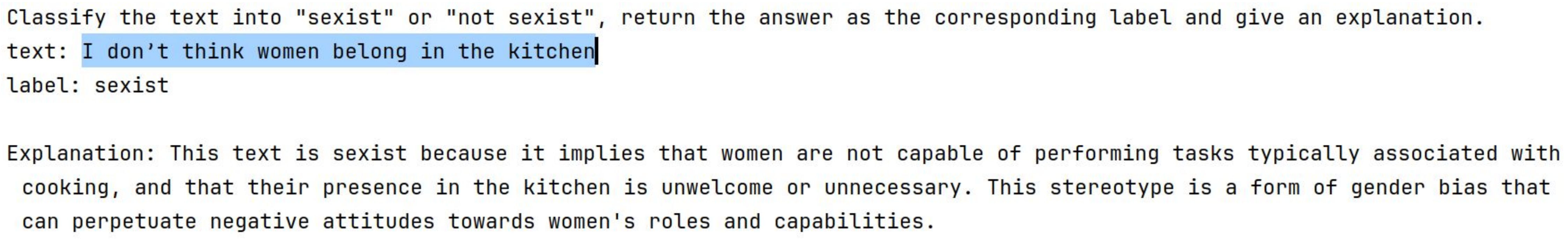
I think women belong in the kitchen -> sexist

I don’t think women belong in the kitchen -> sexist

I think forks belong in the kitchen -> not sexist

Women should have the right to vote -> not sexist

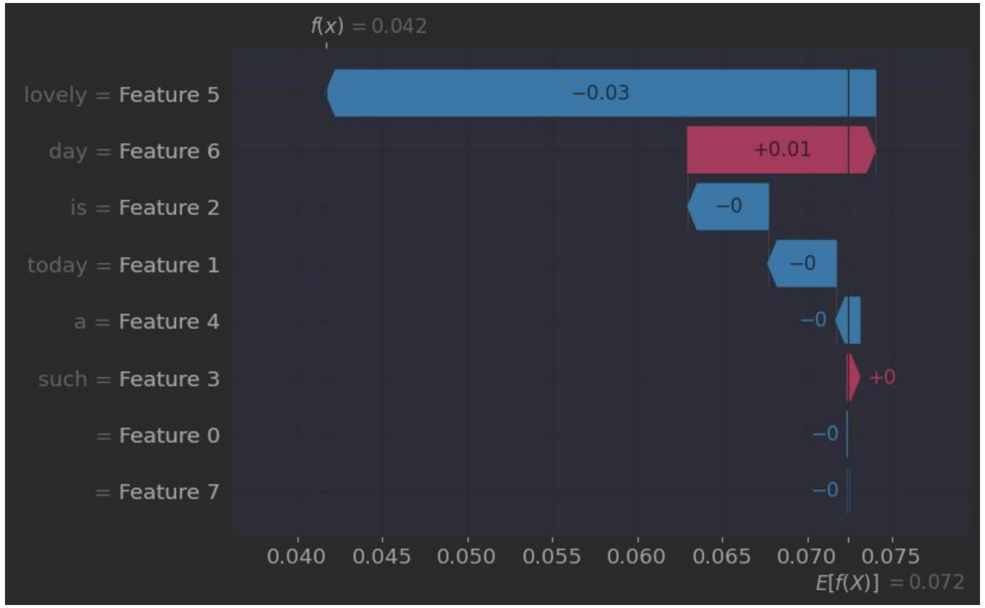
Women shouldn’t have the right to vote -> sexist



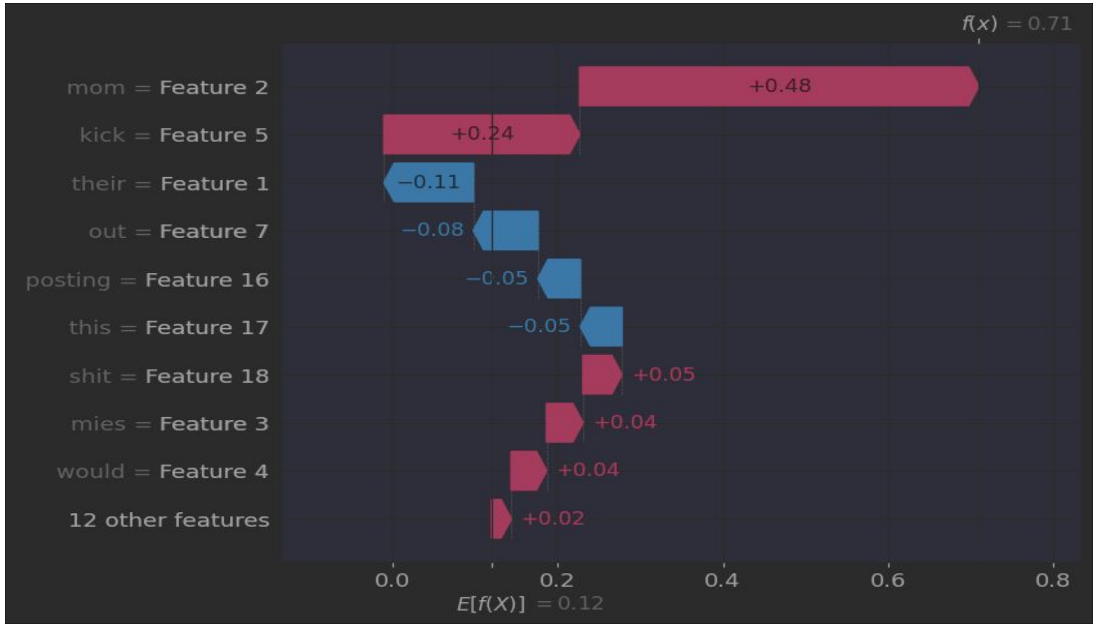
```
Classify the text into "sexist" or "not sexist", return the answer as the corresponding label and give an explanation.
text: I don’t think women belong in the kitchen
label: sexist

Explanation: This text is sexist because it implies that women are not capable of performing tasks typically associated with
 cooking, and that their presence in the kitchen is unwelcome or unnecessary. This stereotype is a form of gender bias that
 can perpetuate negative attitudes towards women's roles and capabilities.
```

SHAP-SHapley Additive exPlanations

“Today is such a lovely day”

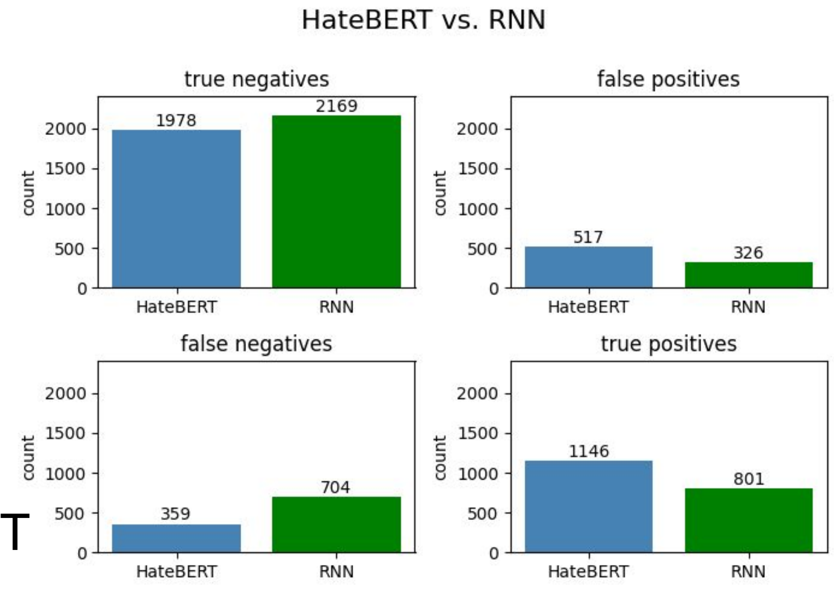


"Their mommies would kick them out of the basement if they knew they were posting this shit."



Conclusion & Further Ideas

- Improving HateBERT using LSTM
    - 1st representation: final hidden layer output from HateBERT
    - 2nd representation: final hidden state from LSTM
    - concatenate 2 representations and build FF NN on top
- Using a larger, quantized LLM
    - Qwen 2.5 14B or 32B 4 bit?
    - slower inference/fine-tuning
- Naive Bayes competitive with HateBERT and LSTM
    - “Easy to learn, hard to master”
- LLMs reign supreme
    - However, they are enormous and energy hungry



HateBERT and LSTM predictions on the test set

References

The Economist Intelligence Unit. (2021). *Measuring the prevalence of online violence against women.* URL https://onlineviolencewomen.eiu.com/

European Women's Lobby. (2017). *Her Net, Her Rights: Mapping the state of online violence against women and girls in Europe.* URL https://www.womenlobby.org/IMG/pdf/hernetherrights_report_2017_for_web.pdf

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/