TECHNISCHE
UNIVERSITÄT
WIEN

Detection of Online Sexism

Babak Bayani (12347302)
Arash Behaein (12324255)
Aditi chauhan (12347667)
Juliane Leibhammer (12408344)

Content

Overview of the task

Objective:

Perform binary classification of short social media utterances to detect sexism (sexist/not sexist)

Dataset:

- EDOS Dataset (annotated part)

Models Used:

- **RoBERTa**: Robustly optimized BERT for various NLP tasks.
- **Logistic Regression**: Quantitative and Qualitative Results, Tokenization.

Date Preprocessing

- Text Cleaning
  - Lowercasing, removing special characters

- Label Encoding (converting labels to binary format)
- Train-validation-test split (Based on the already provided column in dataset.)

Date Exploration

- Label Distribution
  - Imbalanced Dataset (more non-sexist comments)

- Data analysis
  - Most common words

Word Cloud for Sexist Comments

Model Training - RoBERTa

- **Pre-trained Limitations:** General datasets, lacks task-specific optimization.
- **Fine-Tuning Benefits:**
  - Adapts pre-trained weights for binary classification.
  - Adds a classification head for task-specific predictions.
  - Improves accuracy, precision, and recall on custom labels.
- **Steps for Fine-Tuning:**

1. Pre-process and tokenize data for RoBERTa.
2. Add classification head for binary output.
3. Train model using the **training set (DF_train)** and validate on **development set (DF_dev)**.
4. Evaluate on the **test set (DF_test)** with metrics like accuracy, F1 score, and balanced accuracy.

RoBERTa Evaluation

Performance Matrix :

Accuracy	Precision	Recall	F1 Score	Balanced Accuracy	Misclassification Rate
87.38%	74.76%	72.37%	73.55%	82.27%	12.62%

Test Set Classification Report:

Class	Precision	Recall	F1 Score	Support
0 (Sexist)	91%	92%	92%	3030
1 (Non-sexist)	75%	72%	74%	970

Challenges in Classification:

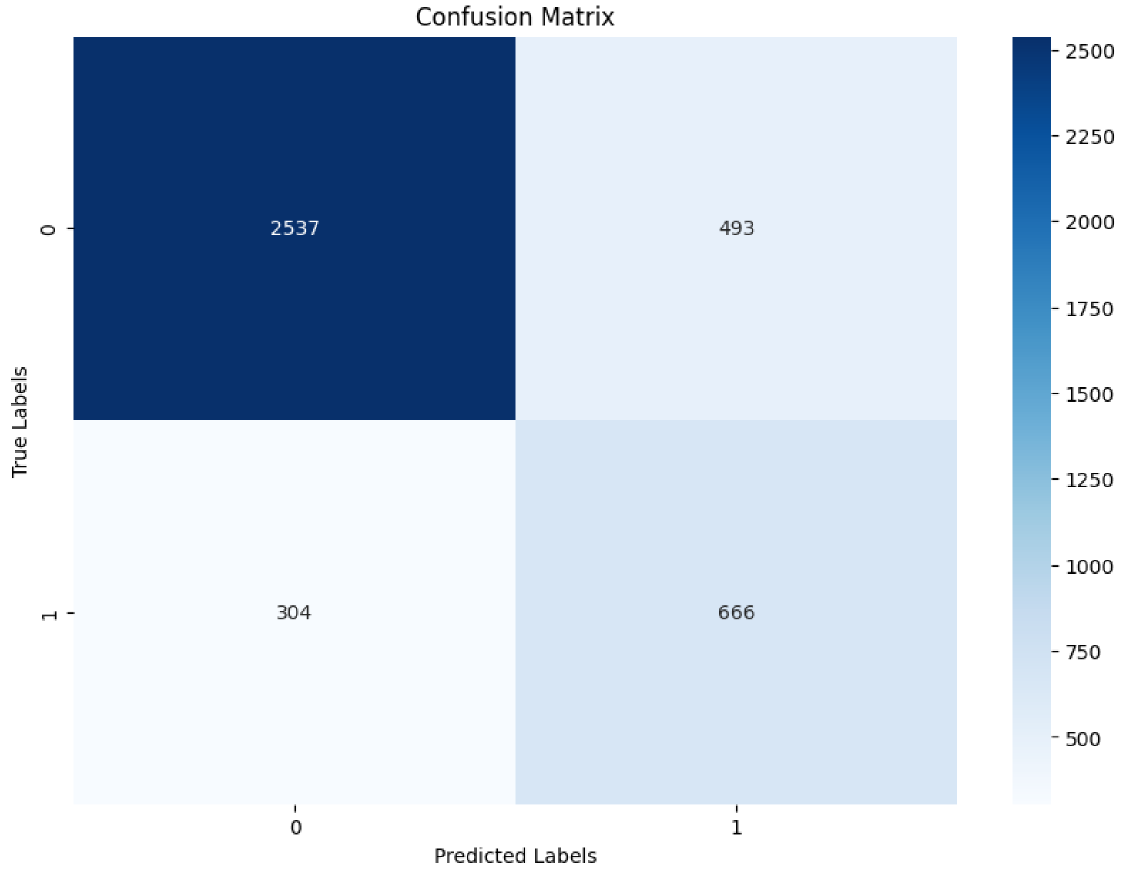
•**Misspellings:** Example: "this bicth should be stoped."

•**Context Loss:** Lack of tone or subtleties leads to errors.

•**Bias in Pre-trained Weights:** Requires domain-specific fine-tuning.

Model Training – Logistic Regression

- Accuracy: 0.800
- Balanced Accuracy: 0.762
- Weighted F1: 0.806
- Recall: 0.686
- Precision: 0.575



Logistic Regression - Evaluation

- Text Pattern Analysis
  - most common tokens in sexist samples "bitch", "female", "women“
  - most common tokens in not-sexist samples "women", "like", "girls“
  - "sex" is often in missclassified non-sexist

- Imbalanced data effects performance
  - Model performs well for non-sexist class but struggles with the minority sexist class.
- Tokenization challenges lead to loss of critical context
  - Original Text: most here would fuck her
  - Tokenized Text: ['fuck']

Analysis

- When analyzing the results we had problems with the given labeling of the dataset

→ Guidelines for Sexism detection

→ Relabel subset of dataset

Guidelines for Sexism

- **Gender-Based Stereotyping and Generalizations**
  Sexist Example: "Women are too emotional to be leaders."

- **Derogatory Language, Objectification, and Questioning Capabilities**
  Sexist Example: "She only got promoted because of her looks."

- **Gender-Based Gatekeeping and Traditional Gender Roles**
  Sexist Example: "Women don’t belong in tech."

- **Double Standards and Dismissal of Gender-Specific Issues**
  Sexist Example: "He's ambitious, but she's bossy."

Problems of Relabeling

S.no	Text	Original label	Sexist (votes)	Not sexist (votes)
1	years from now shell complain that she was forced into exploiting her sexuality in order to feed herself and sue mankind maga feminismiscancer	1	2	2
2	this is true you should only approach a girl without a greenlight if you dont mind being treated like a homeless man at a car dealership	1	2	2
3	this establishment pussy couldnt fight his way out of his lace panties	1	2	2
4	Go to bankok youll have women sucking in your white boobies	0	2	2

Key words

- Commonly Co-Occurring Words:
  - **Derogatory words:** "bitch", "slut", "whore", "cunt"
  - **Physical descriptors:** "tits", "ass", "fat"
  - **Blame-focused words:** "feminists", "women", "they"
  - **Power-related terms:** "dominant", "submission", "control"

- Slang
  - Thots = that ho over there
  - Smv = sexual market value
  - Tranny = offensive word for transgender person

Key patterns

- **Derogatory Adjectives + Women/Female-Specific Terms**
  - **[insult] + [woman/female noun]**

    "Stupid women", "dirty slut", "clueless white women", "fat commie slut"
  - **[adjective] + [body part]**

    "Nice tits", "big ass", "titless butless hag"

- **Blame or Resentment Language**
  - **[women/they] + [negative trait]**

    "Women are full of shit" "women don’t know how to date"
  - **[blame noun] + women/feminism**

    "White women are to blame" "feminism is cancer"

Key patterns

- **Sexual Objectification**
  - **Action-focused patterns:**
    "Whores out her pussy" "obsess over sex" "fuck her too"
  - **Body-focused patterns:**
    "Nice tits" "fatass who insists her rolls"
  - **Judgments on behavior:**
    "Regret the sex to make it rape" "faithful while I stay a whore"

- **Universal declarations:**
  - "Any woman can get laid no matter what"
  - "All women are attracted to bullies"

Key patterns

- **Dominance-focused patterns:**
  "Masculine dominant presence" "men are more successful"
- **Submission-focused patterns:**
  "Women should submit" "feminists defend second-class citizen ideology"
- **Mocking terms:**
  "Hopeless thots" "lace panties" "shedemon"

Fine-tuning of RoBERTa on relabeld Dataset

- The original **fine_tuned_roberta model** shows strong performance with good accuracy, precision, recall, and F1 scores.
- The **fine_tuned_roberta_v2** with the original test set has a slightly lower accuracy but a higher recall for Class (sexist), which means it better identifies instances of Class (sexist) but at the cost of some of it's precision.
- The **fine_tuned_roberta_v2** with the GPT generated test set has a notably poor performance, especially for Class (sexist), with very low recall and balanced accuracy, indicating it struggles to fit to the GPT-generated data.

RoBERTa vs. Logistic Regression

METRIC	Logistic Regression	RoBERTa (Pre-trained)	RoBERTa (Fine-tuned)	RoBERTa(Chat GPT test set)
Accuracy	80.07%	87.38%	84.05%	55.0%
Precision	57.46%	74.76%	63.43%	100.0%
Recall	68.65%	72.37%	80.82%	10.0%
F1 Score	62.56%	73.55%	71.08%	18.18%
Balanced Accuracy	76.19%	82.27%	82.95%	55.0%
Misclassification	19.93%	12.62%	15.95%	45.0%

Challenges

- Relabeled the dataset of 101 samples
- Created a guideline definition of SEXISM.
- Lack of context, Cultural diversity, Pov as different genders.
- Understanding of slangs - thots, Smv,tranny,dyke.
- Understanding the fine line between Abusive, offensive, racist, psycotic languages or statements.
- Should words like bitch, cunt, pussy, hoe, in a statement make it sexist ?
- Who said the statement (M/F/Q) to whom (M/F/Q) ?
- Praises, Appreciation, Sarcasm, Sugar coated words to be considered sexism ?

Next Steps

- Train the model on a more obvious datasets such as the GPTgenerated one that we used for better specifying the basis of sexism.
- Providing the model with specific words and tags that should be classified as sexist.
  - thots = that ho over there
  - smv = sexuell market value
  - tranny=offensive word for Transgender person
- Including emojies, and defining their meaning.
- Defining abbreviations and acronyms.
- Finding a base line to differentiate the concept of Sexism, Racism and Offensive.

TECHNISCHE UNIVERSITÄT WIEN

Thank you for your attention