Detection of Online Sexism

NLP Project 6

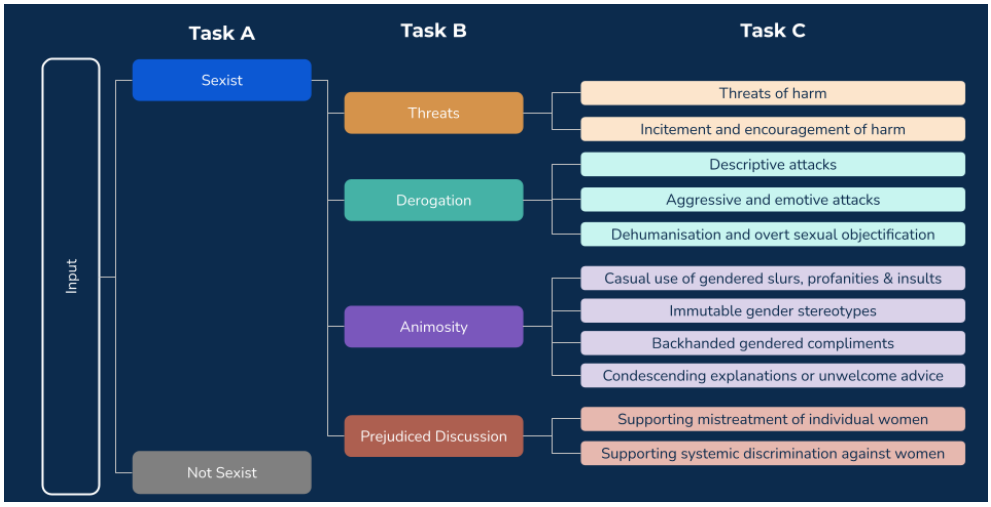
Group 3 Members:

1. Abhliash Datta - 19CS30001
2. Sunanda Mandal - 19CS10060
3. Rohit Raj - 19CS10049
4. Haasita Pinnepu - 19CS30021
5. Vishnu Vardhan - 18CS30022

PROBLEM DESCRIPTION

Online sexism has become a significant challenge. Today, sexist content may be found and evaluated at scale using automated technologies, but most of these tools just classify information into broad categories without providing any extra information. This project supports the development of more accurate and comprehensible English-language sexism detection models by providing fine-grained classifications for sexist content.

The project contains three hierarchical subtasks:
TASK A - Binary Sexism Detection
TASK B - Category of Sexism
TASK C - Fine-grained Vector of Sexism



DATASET

The dataset has 4 columns, the text, it's label_sexist, the label_category and the label_vector

	text	label_sexist	label_category	label_vector
0	"Peaceful Muslim Migrants" gang rape 26 year o...	not sexist	none	none
1	Man watching this was so hard. I dont know how...	not sexist	none	none
2	My mother is not a slut and never was a slut. ...	sexist	3. animosity	3.1 casual use of gendered slurs, profanities,...
3	This is so embarrassing. A. he's hitting on so...	sexist	2. derogation	2.2 aggressive and emotive attacks
4	This. over 80% of men are "betas" or "low-qual...	sexist	3. animosity	3.2 immutable gender differences and gender st...
...	...	...	...	...
7995	If you're not attracted to her, don't go for her?	not sexist	none	none
7996	Also it's not "white nighting" to go try to ta...	not sexist	none	none
7997	There is nothing to destroy.	not sexist	none	none
7998	just demonstrates the wild lack of agency wome...	sexist	2. derogation	2.1 descriptive attacks
7999	There will be no worshiping of women.	not sexist	none	none

8000 rows × 4 columns

This is the training Dataset.

Hyper-Parameter Tuning

For all tasks, we used grid search for hyper-parameter tuning to find optimal learning rate, weight decay rate and number of epochs and ran experiments based on the best results obtained during experiments.

The set of tried hyper parameters are as shown in the table:

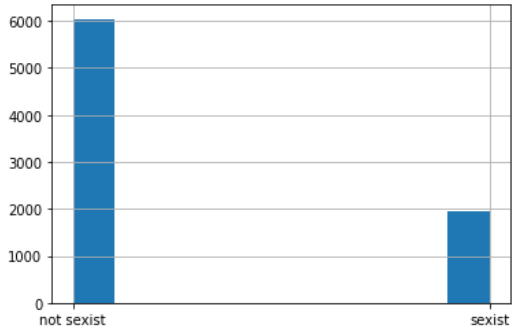
Hyper Parameter	Value1	Value2	Value3
Learning Rate	1e-5	2e-5	5e-5
Weight Decay Rate	0.1	0.01	0.05
Number of Epochs	3	5	10

For every task, we chose the best performing set (i.e. giving highest accuracy) of hyper parameters (Accuracy for other sets has not been indicated to very large search space).

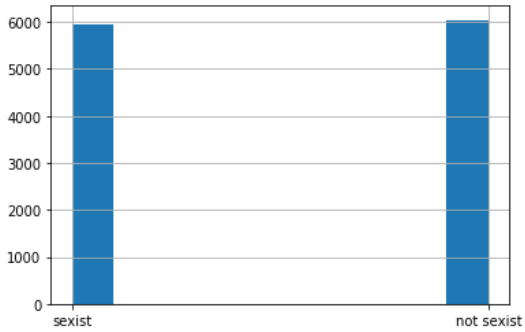
TASK A - Binary Sexism Detection

A two-class (or binary) classification where systems have to predict whether a post is sexist or not sexist.

Class Imbalance: We observe there is significant class imbalance in favor of “not sexist” label, as shown below.



We attempt to decrease the class imbalance by random over-sampling arriving at:



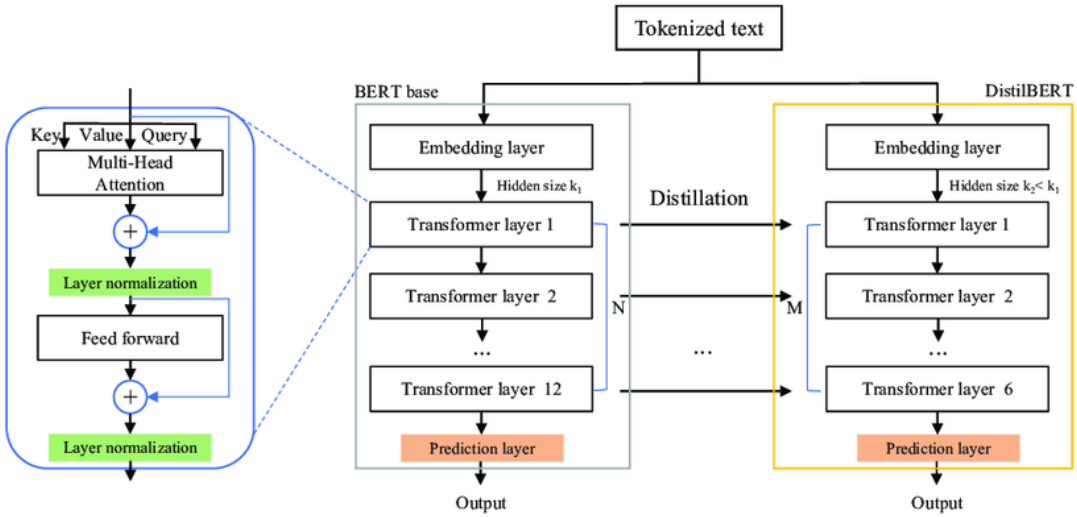
Model Description: We used “DistilBert-Uncased” for the classification task. The configuration of model used is:

```
DistilBertConfig {

"_name_or_path": "distilbert-base-uncased-finetuned-sst-2-english",

"activation": "gelu",

"architectures": [

"DistilBertForSequenceClassification"

],
```

```
"attention_dropout": 0.1,

"dim": 768,

"dropout": 0.1,

"finetuning_task": "sst-2",

"hidden_dim": 3072,

"id2label": {

  "0": "NEGATIVE",

  "1": "POSITIVE"

},

"initializer_range": 0.02,

"label2id": {

  "NEGATIVE": 0,

  "POSITIVE": 1

},

"max_position_embeddings": 512,

"model_type": "distilbert",

"n_heads": 12,

"n_layers": 6,

"output_past": true,

"pad_token_id": 0,

"qa_dropout": 0.1,

"seq_classif_dropout": 0.2,

"sinusoidal_pos_embds": false,

"tie_weights_": true,

"transformers_version": "4.24.0",
```

```
    "vocab_size": 30522
}
```

Model Architecture: The architecture of DistilBert is as shown below:



Hyperparameters:

```
Num examples = 8000
Num Epochs = 10
Instantaneous batch size per device = 16
Total train batch size (w. parallel, distributed & accumulation) = 16
Gradient Accumulation steps = 1
Total optimization steps = 5000
```

Results:

The Classification Report, F1 Score and Accuracy of the model performance on the test Dataset are

```
              precision    recall  f1-score   support

           0       0.89      0.92      0.90      1531
           1       0.71      0.61      0.66       469

    accuracy                           0.85      2000
   macro avg       0.80      0.77      0.78      2000
weighted avg       0.84      0.85      0.85      2000

F1_score = 0.6567505720823799
Accuracy_score = 0.85
```

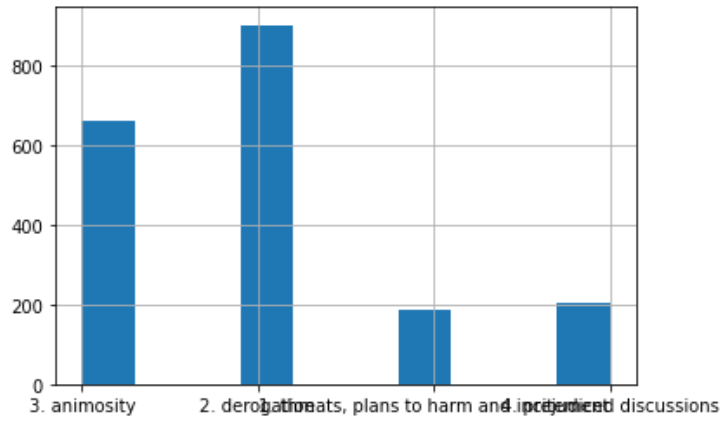
Explainability:

In our attempt to explain output of model, we used transformer_interpret library to check the words which are given most of weights during classification. Our observations are:

- Model mainly focused on generic words like “Women”, “Wife”, “Partner”, etc for classifying text into category - 0
- For classification in category - 1, words relating of sexual orientation like “Lesbian”, “Poly”, etc were given high weights

TASK B: Category of Sexism

For posts which are sexist, a four-class classification where systems have to predict one of four categories: (1) threats, (2) derogation, (3) animosity, (4) prejudiced discussion.



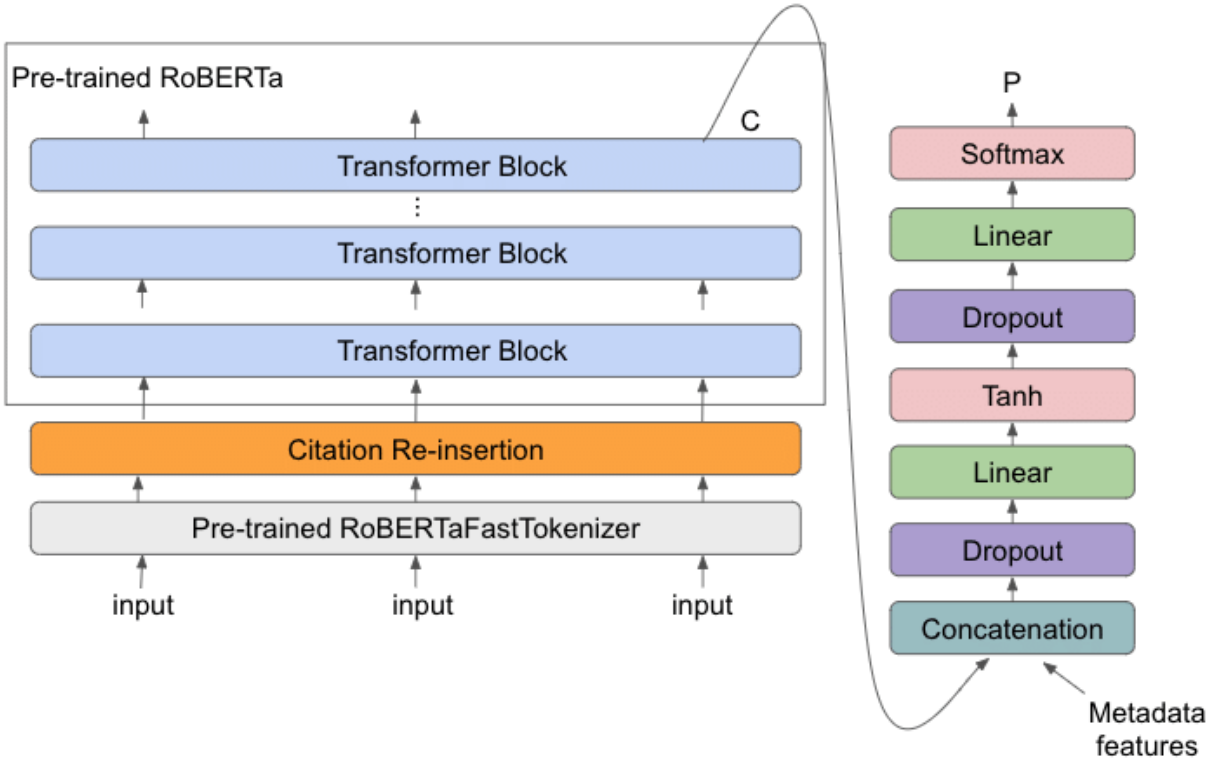
The classification labels in the dataset are: '1. threats, plans to harm and incitement', '2. Derogation', '3. animosity', '4. prejudiced discussions'.

Model Description: Used the RoBERTa base model. The configuration of the model is as shown below:

```
Model config RobertaConfig {
  "_name_or_path": "roberta-base",
  "architectures": [
    "RobertaForMaskedLM"
  ],
  "layer_norm_eps": 1e-05,
  "max_position_embeddings": 514,
  "model_type": "roberta",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "pad_token_id": 1,
  "position_embedding_type": "absolute",
  "transformers_version": "4.24.0",
  "type_vocab_size": 1,
```

```
  "use_cache": true,
  "vocab_size": 50265
}
```


Model Architecture: The architecture of RoBERTa is as shown below:



Hyperparameters:

```
Num examples = 1959
Num Epochs = 3
Instantaneous batch size per device = 16
Total train batch size (w. parallel, distributed & accumulation) = 16
Gradient Accumulation steps = 1
Total optimization steps = 369
```

Results:

The Classification Report, F1 Score and Accuracy of the model performance on the test Dataset are

```
              precision    recall  f1-score   support

           0       0.62      0.64      0.63        33
           1       0.60      0.70      0.64       218
           2       0.58      0.44      0.50       171
           3       0.57      0.57      0.57        47

    accuracy                           0.59       469
   macro avg       0.59      0.59      0.59       469
weighted avg       0.59      0.59      0.58       469

F1_score = 0.5846283306323755
Accuracy_score = 0.5906183368869936
```

Explainability:

In our attempt to explain output of model, we used transformer_interpret library to check the words which are given most of weights during classification. Our observations are:

- Model mainly focused on generic words like “Women”, “Man”, “Name”, etc for classifying text into category - 0
- For classification in category - 1, words relating of sexual orientation like “Lesbian”, “Poly”, etc were given high weights but most of the focused words were generic
- For classification in category - 2, salngs like “Balls”, “Slut”, etc were given high weights
- For classification in category - 1, abusive words llike “Fuck”, “Ass”, etc were given high weights

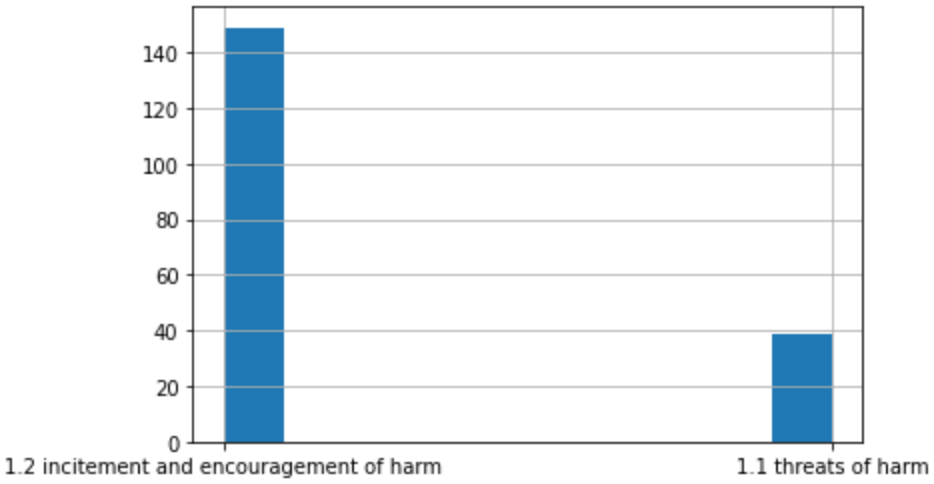
TASK C: Fine-Grained Vector of Sexism

For posts which are sexist, an 11-class classification where systems have to predict one of 11 fine-grained vectors.

We now classify each of the 4 categories of sexism into their respective subcategories.

PART 1:

The label Category is '1. threats, plans to harm and incitement', and these are to be further classified into '1.1 threats of harm', '1.2 incitement and encouragement of harm'.



Model Description: Used the RoBERTa base model. The configuration and architecture of the model is similar to the one used in Task B.

Hyperparameters:

```
Num examples = 188

Num Epochs = 5

Instantaneous batch size per device = 16

Total train batch size (w. parallel, distributed & accumulation) = 16

Gradient Accumulation steps = 1

Total optimization steps = 60
```

Results:

The Classification Report, F1 Score and Accuracy of the model performance on the test Dataset are

```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         6
           1       0.82      1.00      0.90        27

    accuracy                           0.82        33
   macro avg       0.41      0.50      0.45        33
weighted avg       0.67      0.82      0.74        33

F1_score = 0.7363636363636363
Accuracy_score = 0.8181818181818182
```

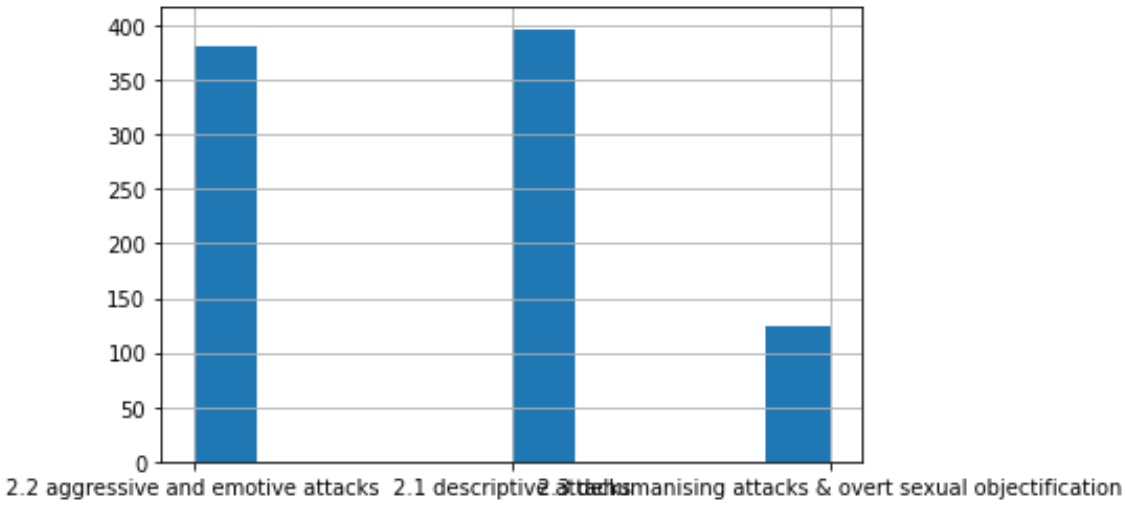
Explainability:

In our attempt to explain output of model, we used transformer_interpret library to check the words which are given most of weights during classification. Our observations are:

- Model mainly focused on generic words like “Streets”, “Name”, etc for classifying text into category - 0 but slangs against women like “Bitch” and words related to violence like “Harm”, “Hitting”, etc were also given high weights
- For classification in category - 1, most words were generic but words corresponding to violence like “Kick”, “Punishment”, etc were given high weightage

PART 2:

The label Category is ‘2. derogation’, and these are to be further classified into '2.1 descriptive attacks', '2.2 aggressive and emotive attacks', '2.3 dehumanising attacks & overt sexual objectification'.



Model Description: Used the RoBERTa base model. The Architecture and architecture of the model is similar to the one used in Task B.

Hyperparameters:

```
Num examples = 903

Num Epochs = 5

Instantaneous batch size per device = 16

Total train batch size (w. parallel, distributed & accumulation) = 16

Gradient Accumulation steps = 1

Total optimization steps = 285
```

Results:

The Classification Report, F1 Score and Accuracy of the model performance on the test Dataset are:

```
              precision    recall  f1-score   support

           0       0.86      0.78      0.82       112
           1       0.75      0.82      0.78        89
           2       0.35      0.41      0.38        17

    accuracy                           0.77       218
   macro avg       0.65      0.67      0.66       218
weighted avg       0.78      0.77      0.77       218

F1_score = 0.7696587395994253
Accuracy_score = 0.7660550458715596
```

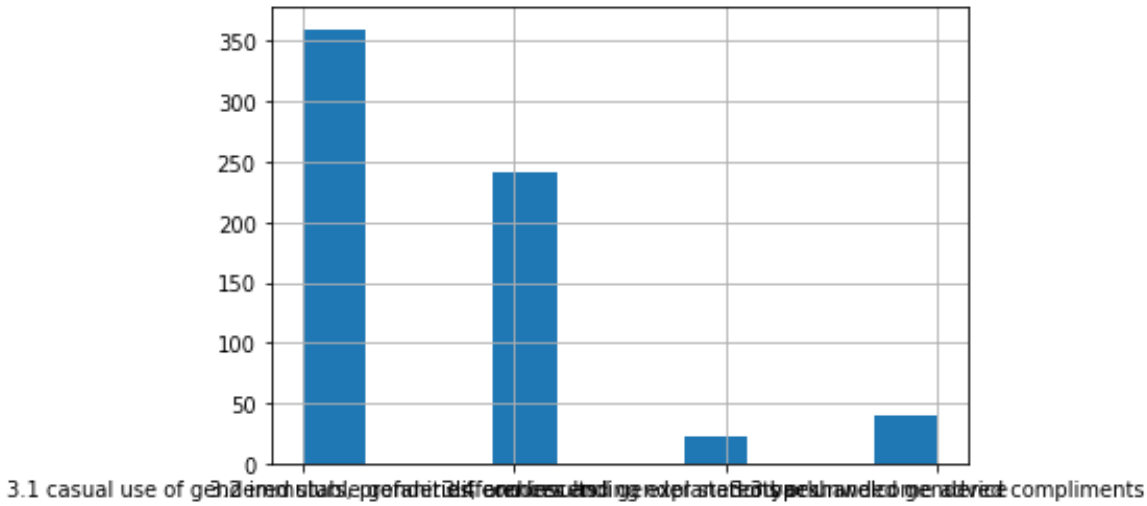
Explainability:

In our attempt to explain output of model, we used transformer_interpret library to check the words which are given most of weights during classification. Our observations are:

- Model mainly focused on generic words for classifying text into category - 0 but few words of interest included “Allegations”, “Block”, “Power”, etc
- For classification in category - 1, words relating of sexual orientation like “Lesbian”, etc and slangs like “Bitch” were given high weights
- For classification in category - 2, words of interest included “Shoot”, “Sex”, “Pigs”, etc

PART 3:

The label Category is ‘3. animosity’, and these are to be further classified into '3.1 casual use of gendered slurs, profanities, and insults', '3.2 immutable gender differences and gender stereotypes', '3.3 backhanded gendered compliments', '3.4 condescending explanations or unwelcome advice'.



Model Description: Used the RoBERTa base model. The configuration and architecture of the model is similar to the one used in Task B.

Hyperparameters:

```
Num examples = 903

Num Epochs = 5

Instantaneous batch size per device = 16

Total train batch size (w. parallel, distributed & accumulation) = 16

Gradient Accumulation steps = 1

Total optimization steps = 285
```

Results:

The Classification Report, F1 Score and Accuracy of the model performance on the test Dataset are

```
              precision    recall  f1-score   support

           0       0.83      0.77      0.80       112
           1       0.73      0.80      0.76        89
           2       0.33      0.35      0.34        17

    accuracy                           0.75       218
   macro avg       0.63      0.64      0.64       218
weighted avg       0.75      0.75      0.75       218

F1_score = 0.7494257247142717
Accuracy_score = 0.7477064220183486
```

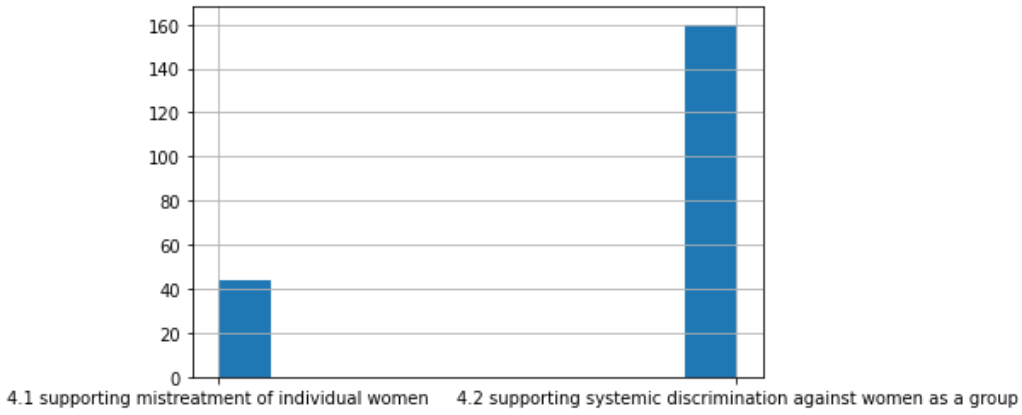
Explainability:

In our attempt to explain output of model, we used transformer_interpret library to check the words which are given most of weights during classification. Our observations are:

- Model mainly focused on generic words for classifying text into category - 0 but few words of interest included “Ass”, “Fuck”, etc
- For classification in category - 1, words of interest included “Clothed”, “Nazi”, “Lesbian”, etc
- For classification in category - 2, words of interest included “Suspicious”, “Pigs”, etc

PART 4:

The label Category is '4. prejudiced discussions', and these are to be further classified into '4.1 supporting mistreatment of individual women', '4.2 supporting systemic discrimination against women as a group'.



Model Description: Used the RoBERTa base model. The configuration and architecture of the model is similar to the one used in Task B.

Hyperparameters:

```
Num examples = 204

Num Epochs = 5

Instantaneous batch size per device = 16

Total train batch size (w. parallel, distributed & accumulation) = 16

Gradient Accumulation steps = 1

Total optimization steps = 65
```

Results:

The Classification Report, F1 Score and Accuracy of the model performance on the test Dataset are

```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        13
           1       0.72      1.00      0.84        34

    accuracy                           0.72        47
   macro avg       0.36      0.50      0.42        47
weighted avg       0.52      0.72      0.61        47

F1_score = 0.6073023377987917
Accuracy_score = 0.723404255319149
```

Explainability:

In our attempt to explain output of model, we used transformer_interpret library to check the words which are given most of weights during classification. Our observations are:

- Model mainly focused on generic words for classifying text into category - 0 but few words of interest included “Rape”, “Arrogant”, etc
- For classification in category - 1, words of interest included “Fuck”, “Accusations”, etc

—----------END—----------