MERNA ABDELBADIE
ZEIN NOUREDDIN

Advanced Machine Learning - Fall 2023

DETECTION OF SEXIST STATEMENTS AGAINST WOMEN

References:

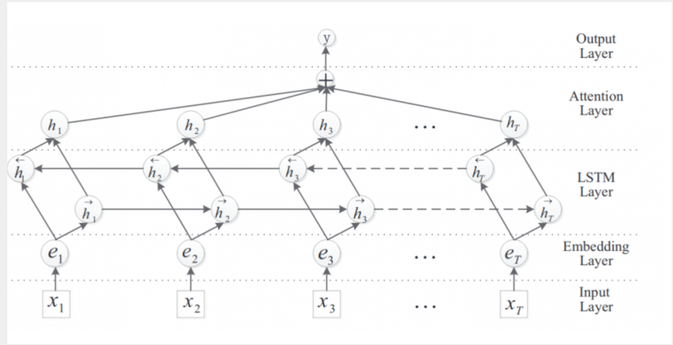
- https://paperswithcode.com/paper/automatic-detection-of-sexist-statements
- https://github.com/rewire-online/edos/blob/main/data/edos_labelled_aggregated.csv
- https://www.kaggle.com/datasets/usharengaraju/dynamically-generated-hate-speech-dataset
- https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech
- https://huggingface.co/datasets/social_bias_frames

INTRODUCTION

The motivation behind addressing this problem lies in the pressing need to create technology that not only recognizes and combats sexism but also contributes to fostering a more equitable and inclusive society, ultimately promoting diversity and fairness across various applications and platforms.

BASELINE MODEL

The baseline model that we experimented with is a bidirectional LSTM model with attention mechanism, with two 128-neuron layers.



DATASET

Our dataset consisted of an augmented version of the concatenation of a processed, balanced version of the EDOS dataset as well as the original dataset used by the model.

UPDATES

- Fine-tuned the hyperparameters
- Fine-tuned the model by pre-training it on general hate speech then training it on sexist speech
- Used bidirectional GRU insted of birdirectional LSTM
- Increased the dataset using data augumentation
- Added user feedback loop that retrains model when it predicts incorrectly

INPUT/OUTPUT

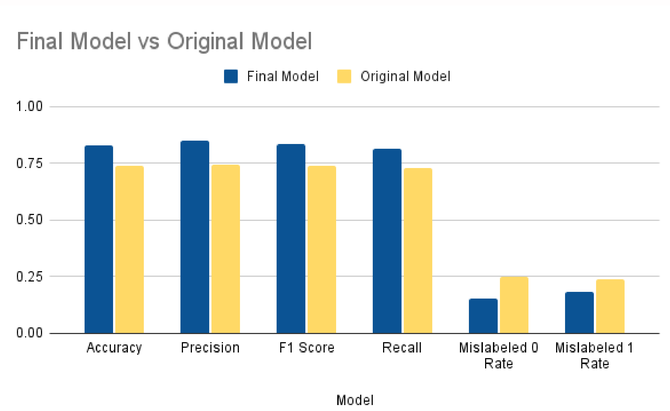
```
Model predictions:
Input: Life is horrible and boring, Prediction: not sexist
Input: Women are not suited for leadership roles., Prediction: sexist
```

CONCLUSION

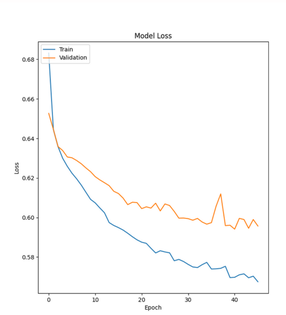
After extensive experimentation with hyperparameters and augmenting the dataset, we successfully enhanced the model's accuracy, elevating it from 0.737 in the original model to 0.8307 in our final model. Moreover, we improved all evaluation metrics compared to the original model, effectively reducing the number of mislabeled statements.

RESULTS

Our optimal model was achieved through the implementation of bidirectional GRU with a dropout rate of 0.5, the Adam regularizer, default learning rate, 32 neurons in the first layer and 64 neurons in the second, a batch size of 64, 100 epochs of training with early stopping, all performed on the augmented dataset.



Best Model



Original Model