Antiviolence NLP

Can a combination of 3 models stop violent language online?

Agenda

1. Scope
2. EDA & Data Processing
3. Model Evaluation
4. Streamlit Presentation
5. Q&A
6. Conclusion

Scope

I believe violence has no place, anywhere, and it should be monitored in the social media community as violent language and cyber mobbing can scar one's self esteem and lead to extreme actions by the victim of such violence.

With this project I want to establish a process which identifies emotions and in case of violent language it reports it automatically to the operating team of the platform. The request is to have the user profile reviewed and see if there's the extreme to ban the profile permanently.

EDA & Data Processing

- 3 data frames: emotion, violence, hate
- data frames modified to only have 2 columns: text, labels
- data frames resampled to have 6000 rows each to improve speed during training
- text cleaned to present only alphanumerical characters
- merged emotional, economic and traditional violence into EET
- stop words removed and text tokenized
- train_test_split with test size=0.2

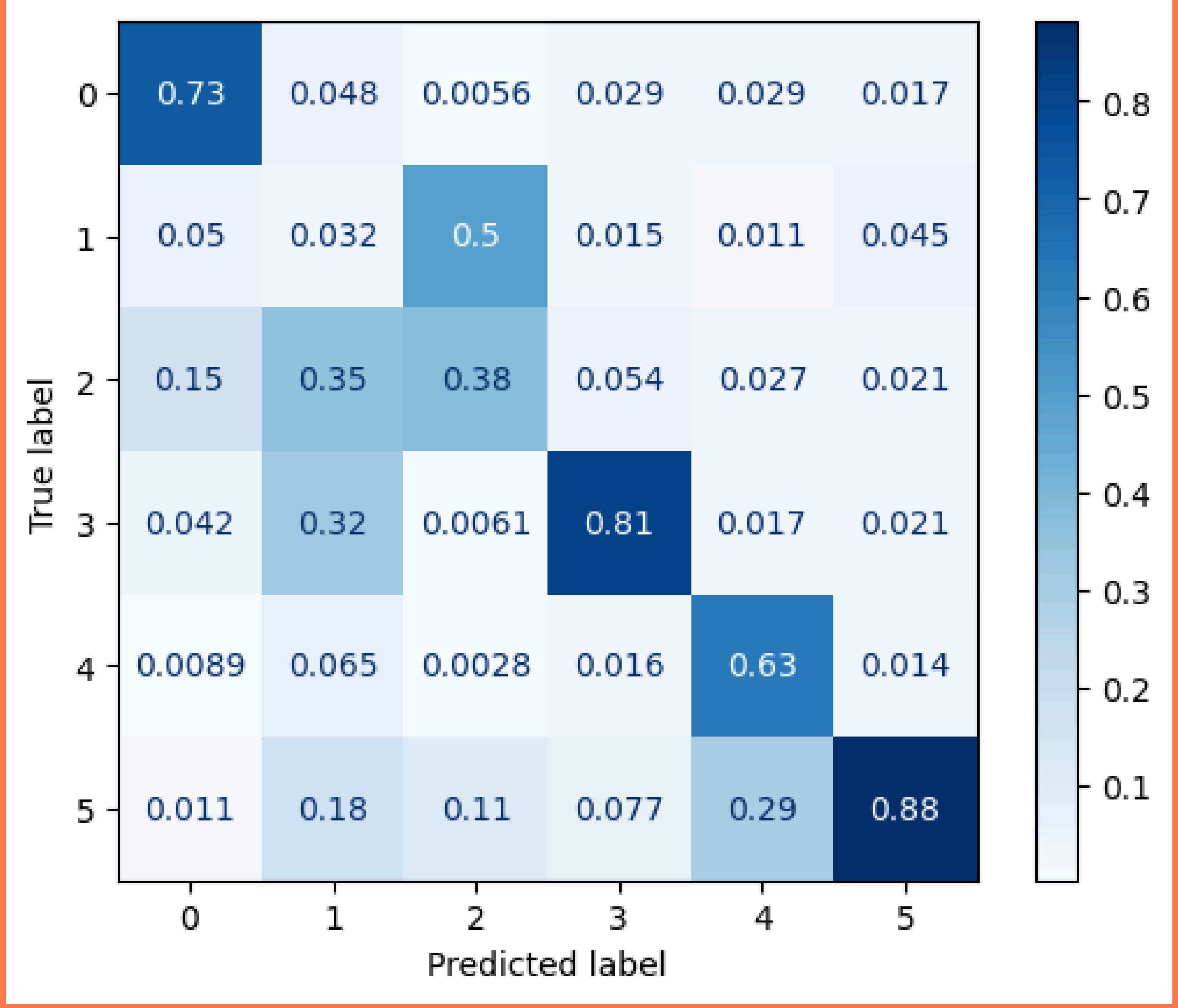
Labels across data frames

- sadness, joy, love, anger, fear and surprise
- Hate, Offensive, Neither
- Sexual Violence, Physical Violence, EET

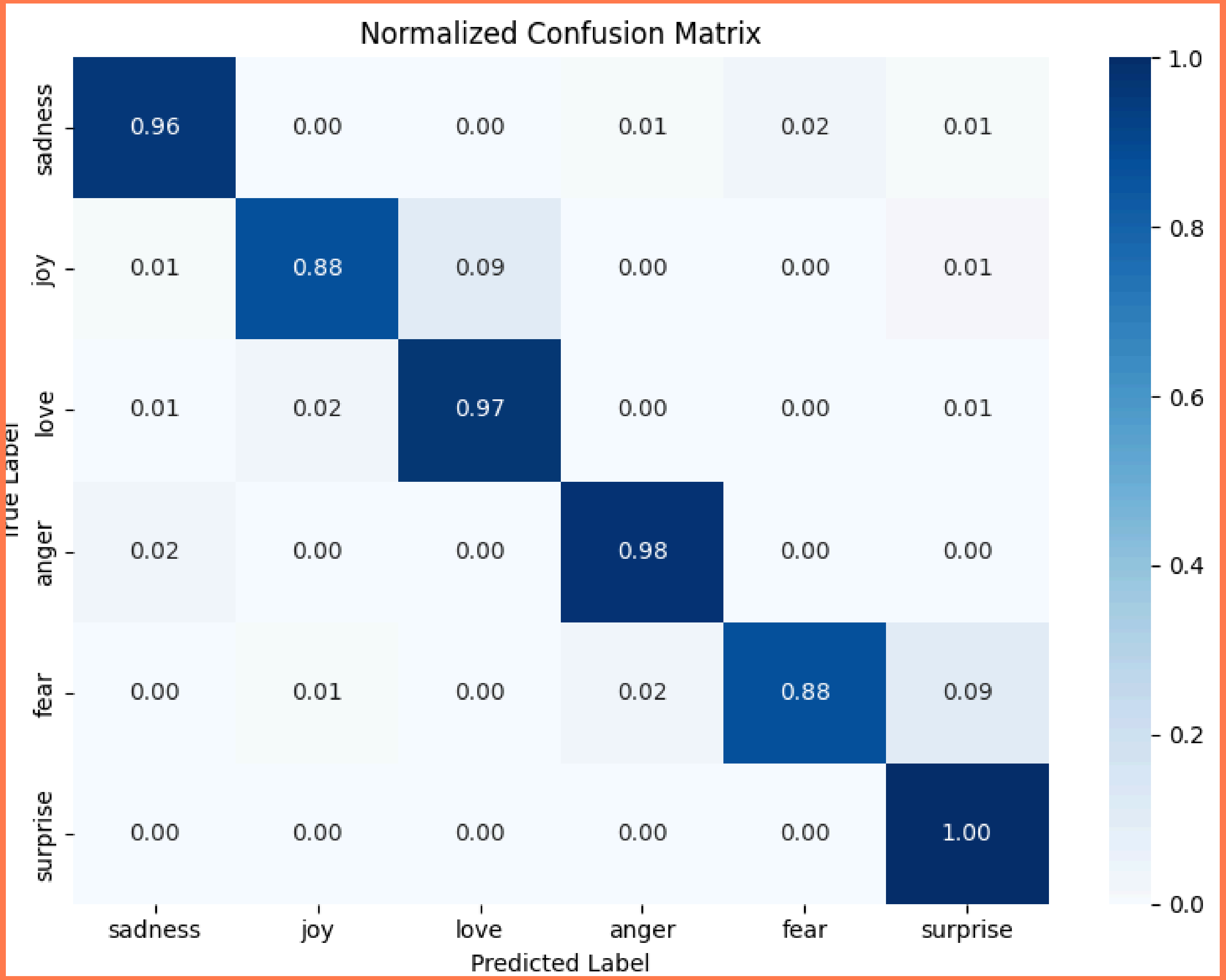
from sklearn.model_selection import utils

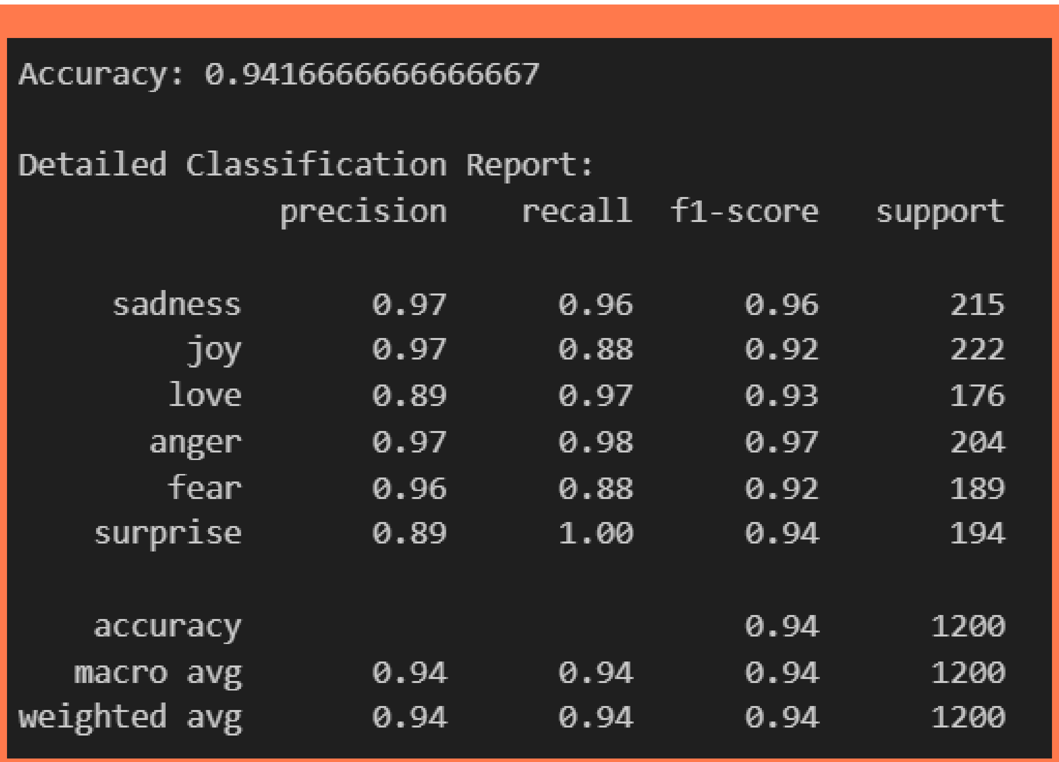
The models presented the best performance after resampling.

Baseline RoBERTa



RoBERTa with Adam optimizer



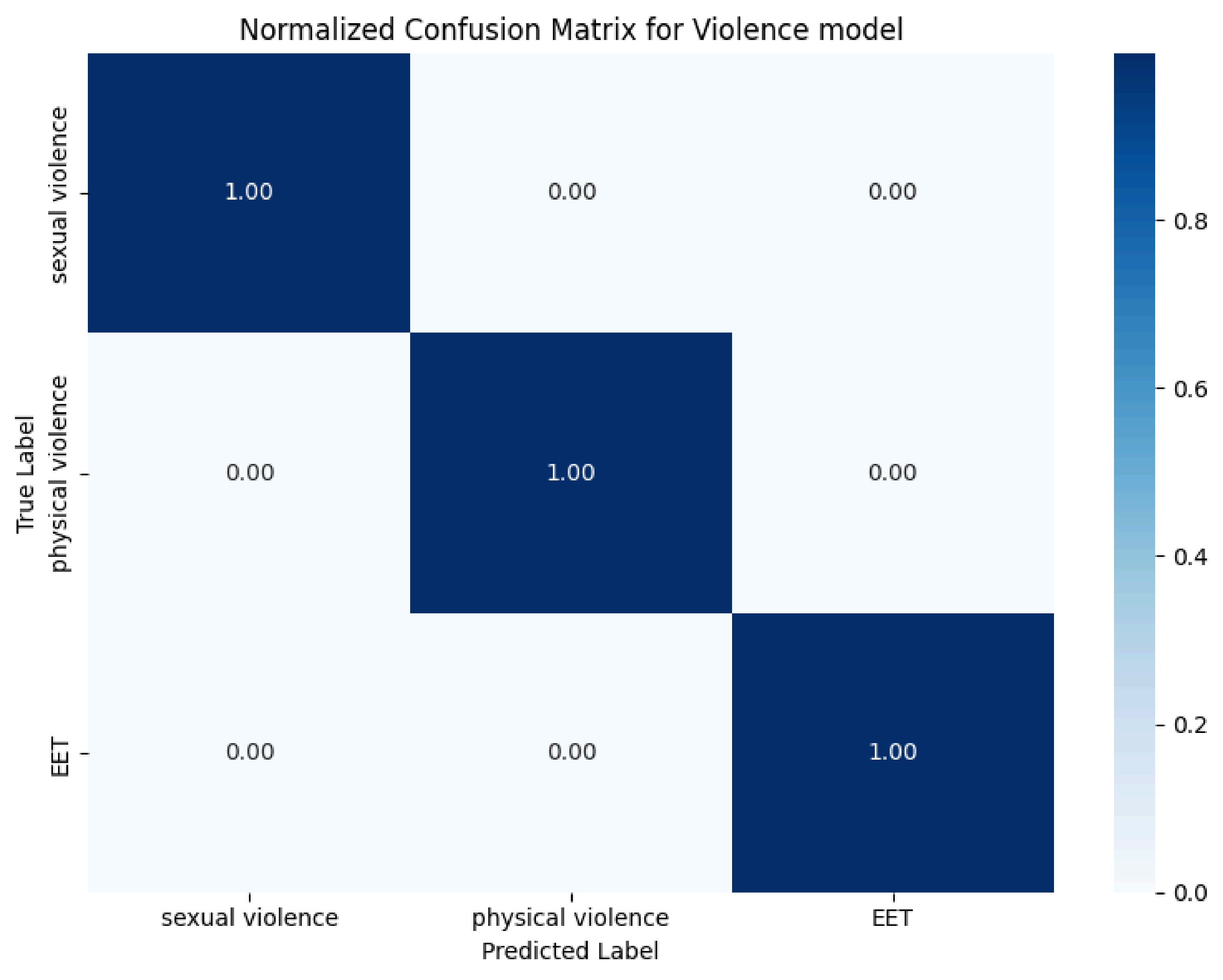


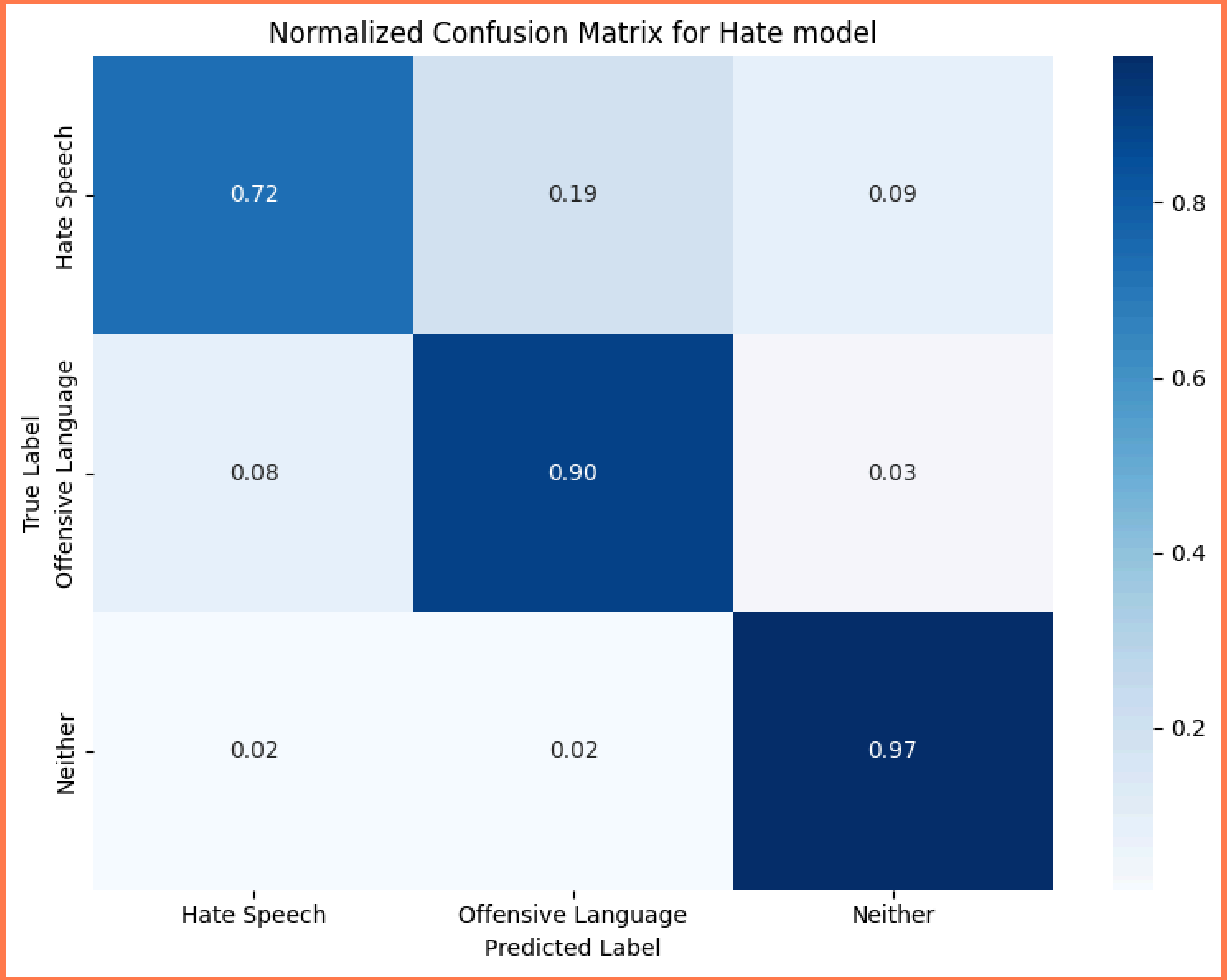
```
Accuracy: 0.9416666666666667

Detailed Classification Report:
              precision    recall  f1-score   support

     sadness       0.97      0.96      0.96       215
         joy       0.97      0.88      0.92       222
        love       0.89      0.97      0.93       176
       anger       0.97      0.98      0.97       204
        fear       0.96      0.88      0.92       189
    surprise       0.89      1.00      0.94       194

    accuracy                           0.94      1200
   macro avg       0.94      0.94      0.94      1200
weighted avg       0.94      0.94      0.94      1200
```

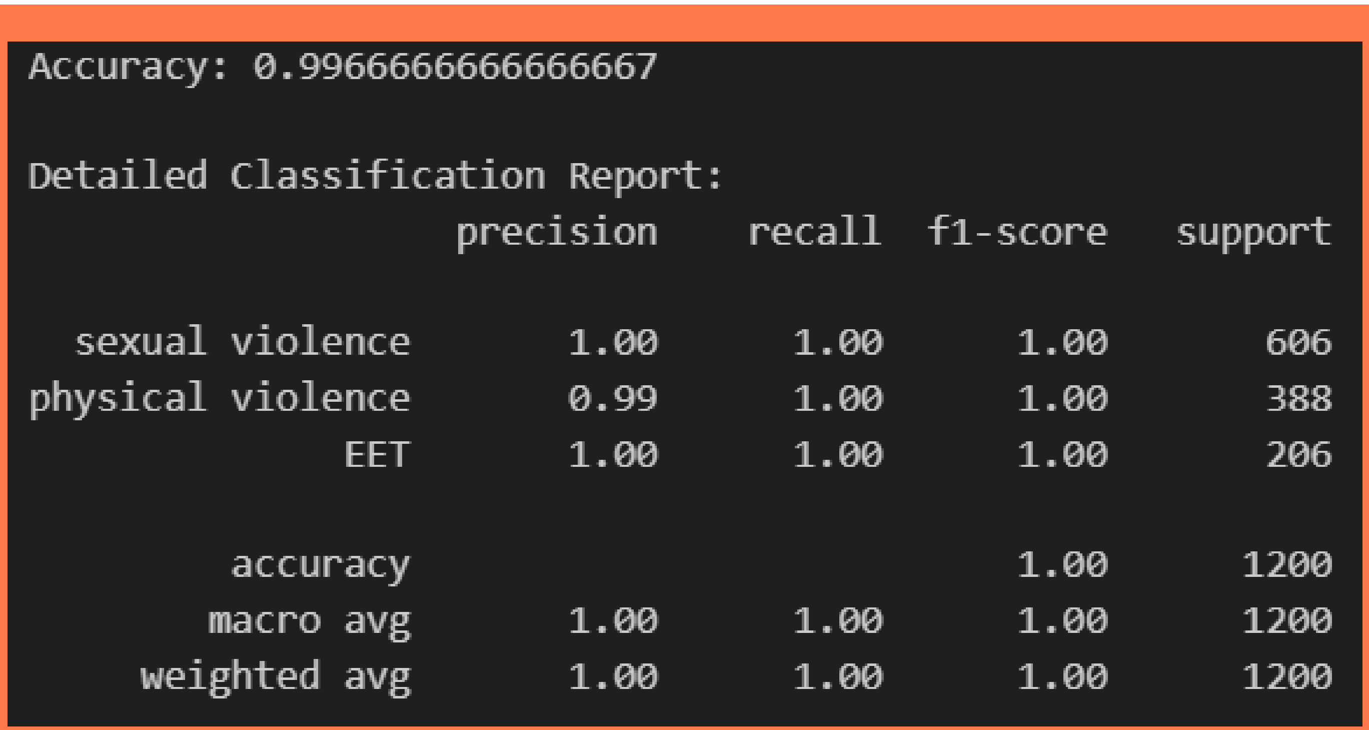
Key Learning for me:
SparseCategoricalCrossentropy really made the different for this project.
The way I understand it, it helps minimizing the loss during training by automatically adjusting the model internal weights.

Violence with Adam optimizer

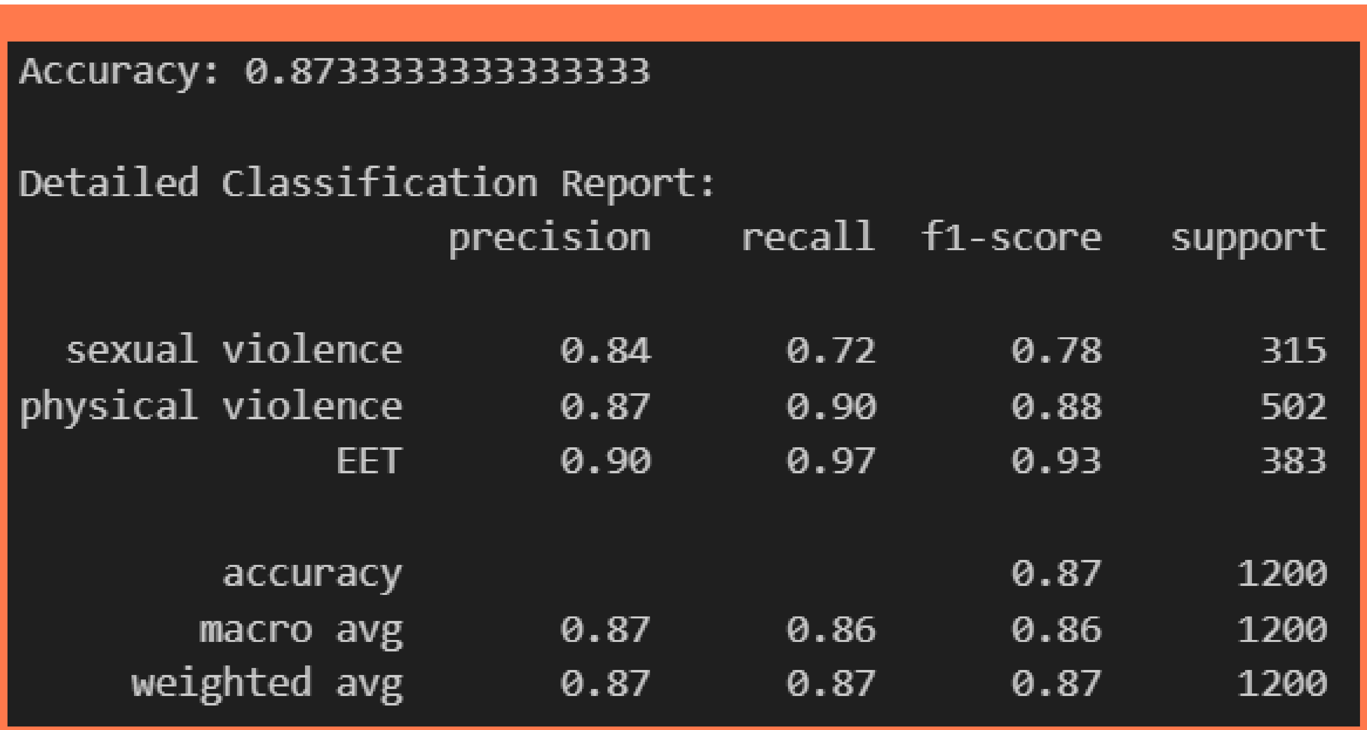
Hate with Adam optimizer







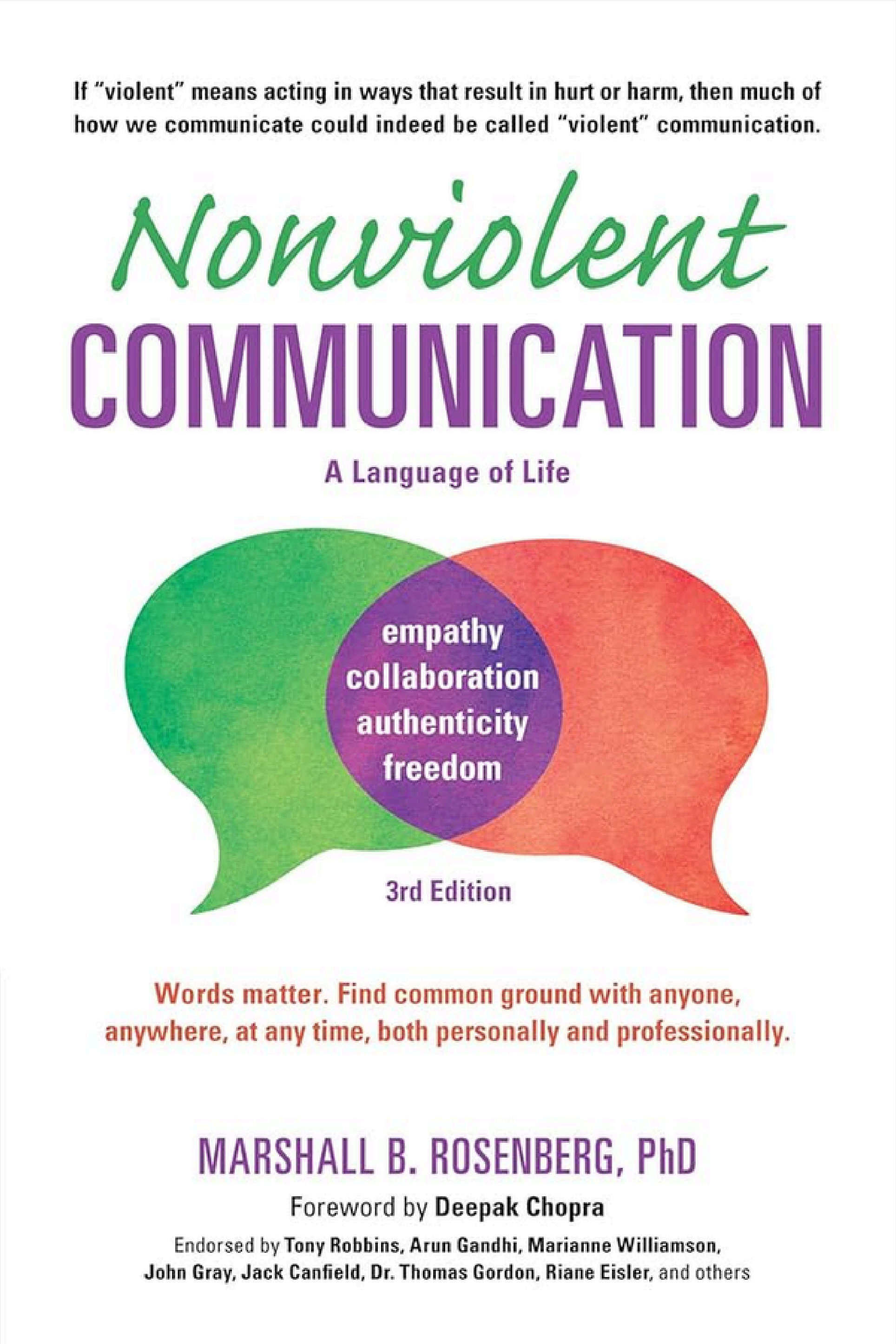
```
Accuracy: 0.9966666666666667

Detailed Classification Report:
                   precision    recall  f1-score   support

  sexual violence       1.00      1.00      1.00       606
physical violence       0.99      1.00      1.00       388
              EET       1.00      1.00      1.00       206

         accuracy                           1.00      1200
        macro avg       1.00      1.00      1.00      1200
     weighted avg       1.00      1.00      1.00      1200
```



```
Accuracy: 0.8733333333333333

Detailed Classification Report:
                   precision    recall  f1-score   support

  sexual violence       0.84      0.72      0.78       315
physical violence       0.87      0.90      0.88       502
              EET       0.90      0.97      0.93       383

         accuracy                           0.87      1200
        macro avg       0.87      0.86      0.86      1200
     weighted avg       0.87      0.87      0.87      1200
```

Q&A

Conclusions



If "violent" means acting in ways that result in hurt or harm, then much of how we communicate could indeed be called "violent" communication.

Nonviolent

COMMUNICATION

A Language of Life

empathy
collaboration
authenticity
freedom

3rd Edition

Words matter. Find common ground with anyone, anywhere, at any time, both personally and professionally.

MARSHALL B. ROSENBERG, PhD

Foreword by **Deepak Chopra**

Endorsed by **Tony Robbins, Arun Gandhi, Marianne Williamson, John Gray, Jack Canfield, Dr. Thomas Gordon, Riane Eisler,** and others