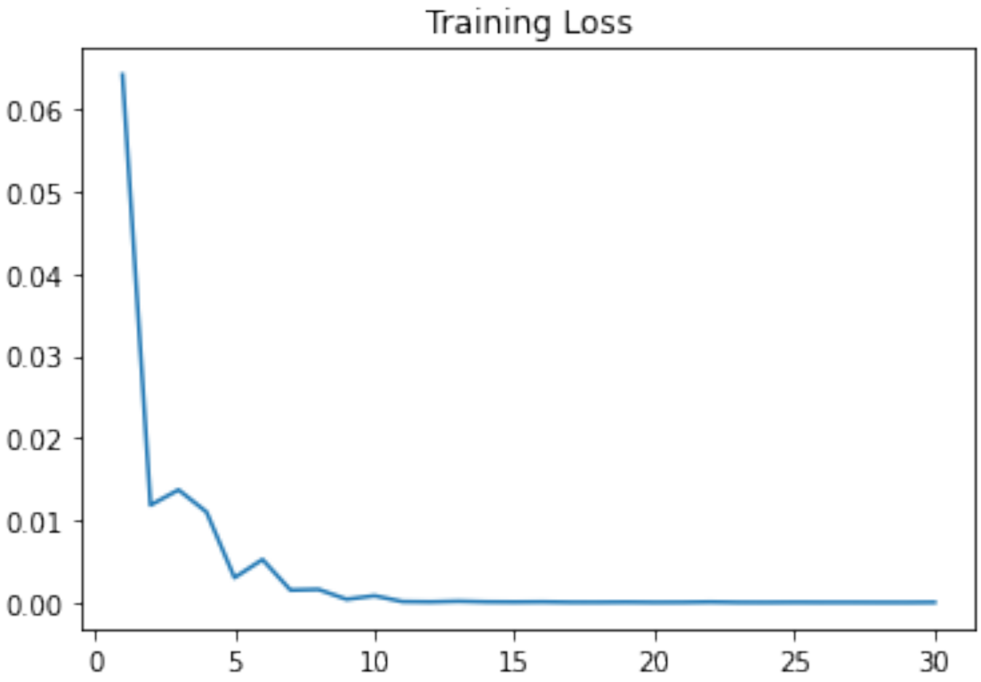
## 1. description of the task and models with references to the original papers and model cards/repository.

I have chosen the Hate Speech Detection problem from the Social Classification tasks. This task is linked to Davidson et al.'s paper called "Automated Hate Speech Detection and the Problem of Offensive Language, " published in 2017. The paper used over 24,802 tweets as its dataset and labeled them into three categories: hate speech, offensive language, or neither. Davidson et al. used logistic regression, naive Bayes, decision trees, random forests, and linear SVMs. Logistic Regression and Linear SVM performed significantly better for them than other models.

I decided to use the Huggingface "hate_speech_offensive" dataset which link was included in the HW1 description and I trained my model as a head on a pre-trained model "bert-base-uncased".

## 2. Learning curve graphs of your training losses from forwarding propagation

My train dataset consists of 80% of the original dataset, 10% evaluation, and 10% test. I used 8 as my batch size (a bigger batch size was causing my computer to crash) and used 30 epochs and at the end, I had 74,370 training steps.



### 3. evaluation metrics used in your experiment

I used accuracy, precision, and F1 scores as my evaluation metrics.

### 4. Test set performance and comparison with score reported in original paper or leaderboard. A justification is needed if it differs from the reported scores.

As I mentioned, I used 10% of my dataset for the test which is around 2479 samples, 10% for validation which also has 2478 samples, and 80% for the train which consists of 19826 samples.

The paper claims that they witnessed 0.91 precision and an F1 score of 0.90 but they also witnessed 40% misclassification of hate speech.

The last model on the leaderboard, “BiLSTM + static BE” which was reported in 2021 reported 80% accuracy, 79% F1 score, and 80% precision. The code of this model is not available on GitHub instead there is another model which is a combination of BiLSTM + Attention + FT but reported only 77% accuracy.

The following numbers are the results from my model:

Validation dataset: Accuracy = 90.71% Precision = 89.84% F1 score = 90.18%

Test dataset: Accuracy = 91.73% Precision = 90.96% F1 score = 91.24%

### 5. training and inference time

I trained my model on Google CoLab Pro with hardware accelerator and high RAM, along with the Huggingface Accelerator library, and Coda device.

I started with 10 epochs which took around 30 minutes and went all the way up to 30 epochs which took 1 hour and 30 minutes.

Evaluation for both validation and test subsets took less than 5 minutes combined.

### 6. Hyperparameters used in your experiment (e.g., number of epochs, learning parameter, dropout rate, hidden size of your model) and other details.

The number of epochs I used is 30 and the learning rate of 1e-5. Using more epochs didn’t make a change in training loss. I also used AdamW as my optimizer.

Bert's model is using 10% dropout and based on my understanding from the model.get_output_embeddings, there are 11 hidden layers in the encoder.

## 7. A minimum of ten incorrectly predicted test samples and their ground-truth labels

I truly apologize for the offensive words from the dataset…

| | Sample | Predicted Label | Ground truth |
|---|---|---|---|
| 1 | &#128520;&#127383; we snap chatted for one night lol. But you\'re cute. Snapchat me back nig | Offensive | Neither |
| 2 | @DavetheBrave1 that hoe on some other shit bro rt and 100x better rn that a nigga high | Offensive | Hate Speech |
| 3 | Ion kno wat kinda spring clothes imma get my yung nicca dis year we ain\'t doin da polo dis year gotta offset em wit sumthin else | Offensive | Neither |
| 4 | RT @JManziel2: &#8220;@RealZachK0enn: @JManziel2 You're a faggot&#8221;you're a Texas fan... | Offensive | Hate Speech |
| 5 | not scanger | Neither | Offensive |
| 6 | idk you into mouthy spics on welfare with 6 kids and 3 baby daddy\'s? | Offensive | Hate Speech |
| 7 | a majority of my news feed is people arguing about what it means to be a redneck | Neither | Offensive |
| 8 | they made me put on basketball shorts now I look like a confused dyke mosh warrior | Hate Speech | Offensive |
| 9 | Thanks bruh don't ever forget we still on the look out for these faggots | Hate Speech | Offensive |
| 10 | @amck73 And because it sounds like there is NO prenup, he\'s gonna cash in, not her. He apparently has ghetto credit. \'magine that?! | Neither | Offensive |

## 8. potential modeling or representation ideas to improve the errors

The biggest downside was that only 11.6% of the tweets in this dataset are labeled as hate speech and the majority of the tweets are considered to be offensive so there is a huge bias in this dataset. I also noticed that the authors of the tweets, or if they were retweeted, were included in the sentence and can create a bias towards the authors. There were examples where I removed the "@xxxx" from the sample and the model predicted a different label for the sample.

To improve the accuracy, more hate speech samples need to be added to the dataset, and rather than using TF-IDF (which was mentioned in the paper that was their primary way to build their uni/bi/trigrams based on), other methods can be examined as well.

It seems like whenever the model sees some specific words such as b*** or f***, it predicts them as hate speech. Since the trends are constantly changing and these words might change their actual meaning based on the context or where they've been used, for model representation, it is better to consider the whole sentence to find some relations between these offensive words with other words in the same sentence. I think POS tagging or entity detection might help with the model representation.