UNIVERSITY OF TEHRAN

COLLEGE OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NEURAL NETWORK & DEEP LEARNING

CONTEXTUAL EMBEDDING USING HATEBERT AND RNNS

SIAVASH SHAMS

MOHAMMAD HEYDARI

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

UNIVERSITY OF TEHRAN

May. 2022

1 CONTENTS

CONTEXTUAL EMBEDDING + RNNS

1.1

- Remove entities like &lt; &gt; &amp; which gets embedded in the original data.
- In the twitter datasets, there is also other information as retweet, Hashtag, Username and modified tweets. All of this is ignored and removed from the dataset.
- Replace digits with <number>

- Remove Punctuations
- Character normalization (tooooooday -> today)
- Tweets shorter than a length are removed

1.2

We use pretrained Bert model and add LSTM and fully connected layers to it and fine tune the parameters as we do in transfer learning.

From figure1 We can see that the length of tweets are shorter than 40 so we pad all the data to 40 maximum length.

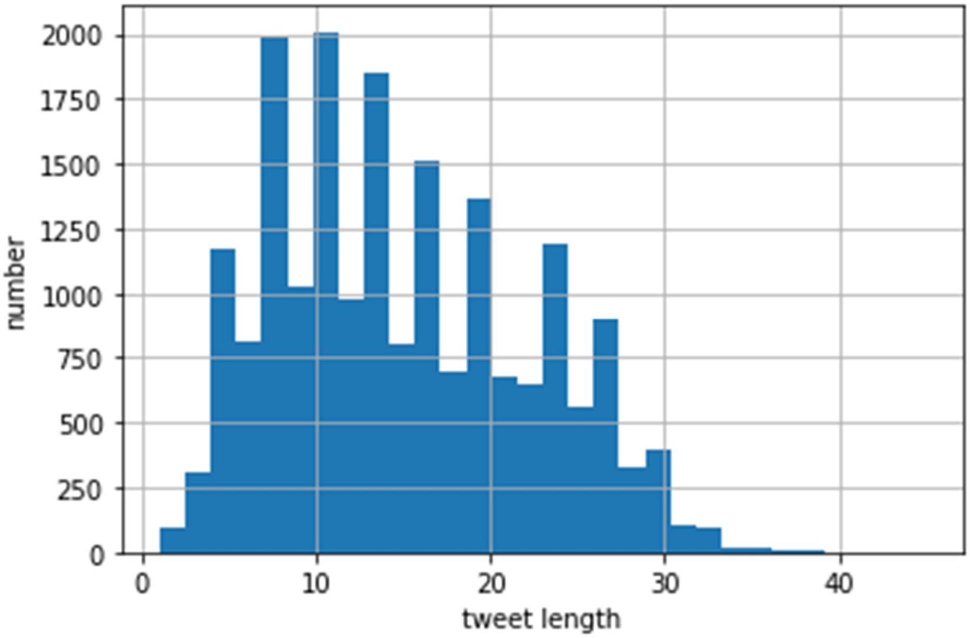


Figure1. Length of the tweets

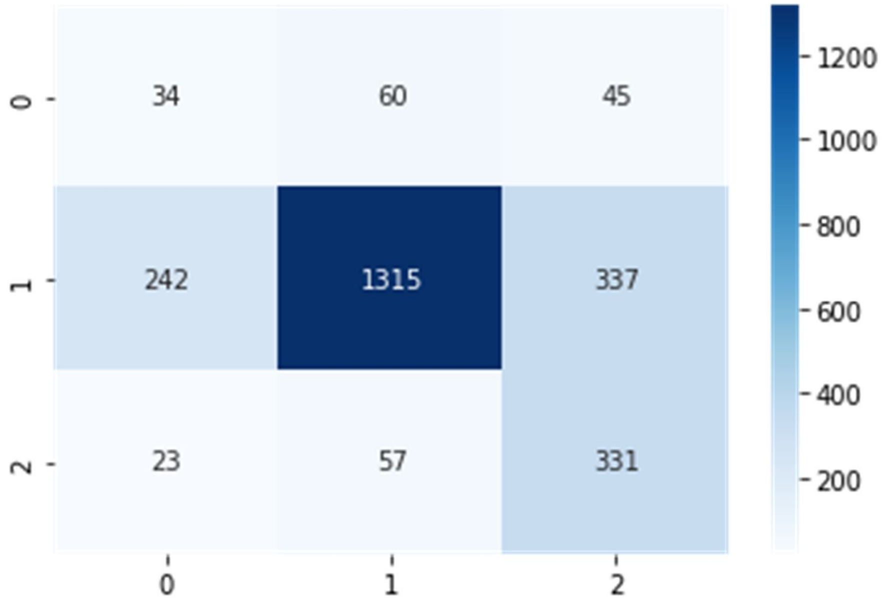


Figure2. Prediction of the context of the tweets (0 means hate, 1 means offensive, 2 means neutral)

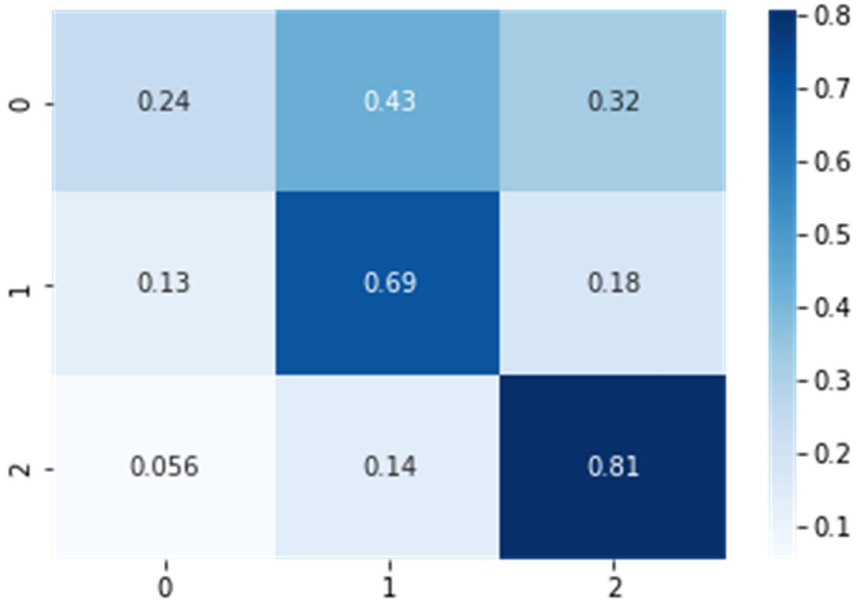


Figure3. Prediction of the context of the tweets (0 means hate, 1 means offensive, 2 means neutral)

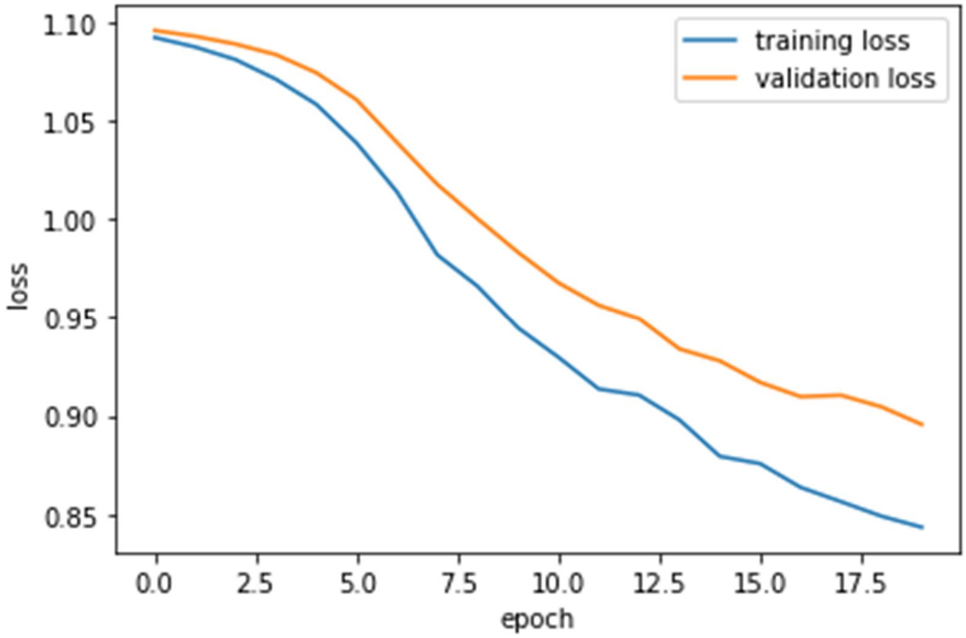
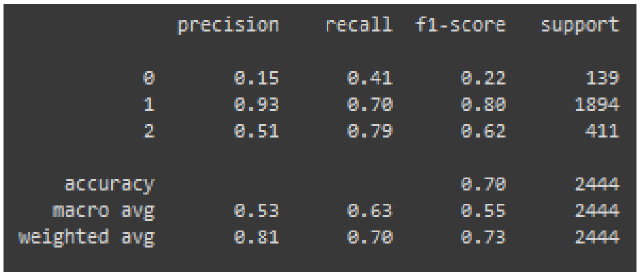


Figure4. Training loss vs validation loss



```
              precision    recall  f1-score   support

           0       0.15      0.41      0.22       139
           1       0.93      0.70      0.80      1894
           2       0.51      0.79      0.62       411

    accuracy                           0.70      2444
   macro avg       0.53      0.63      0.55      2444
weighted avg       0.81      0.70      0.73      2444
```

1.3

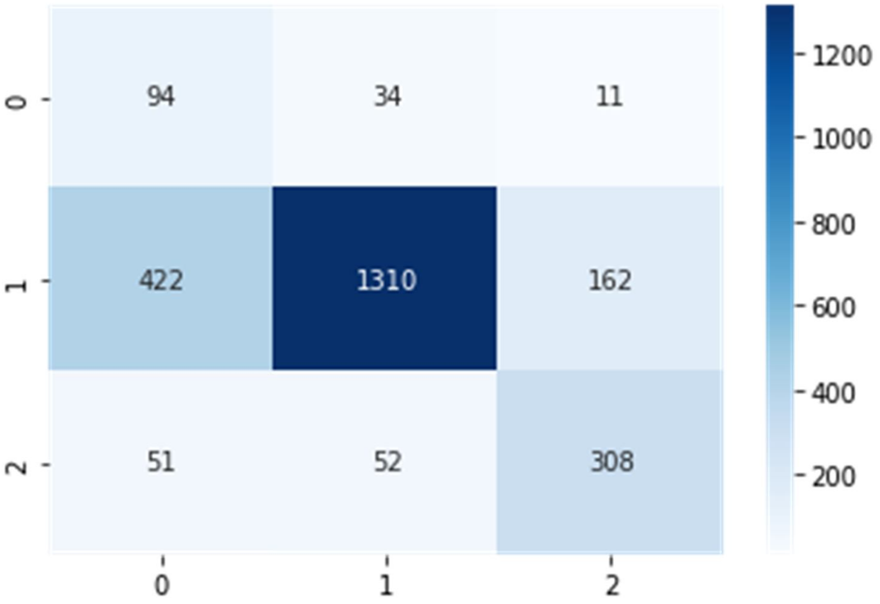


Figure5. Prediction of the context of the tweets (0 means hate, 1 means offensive, 2 means neutral)

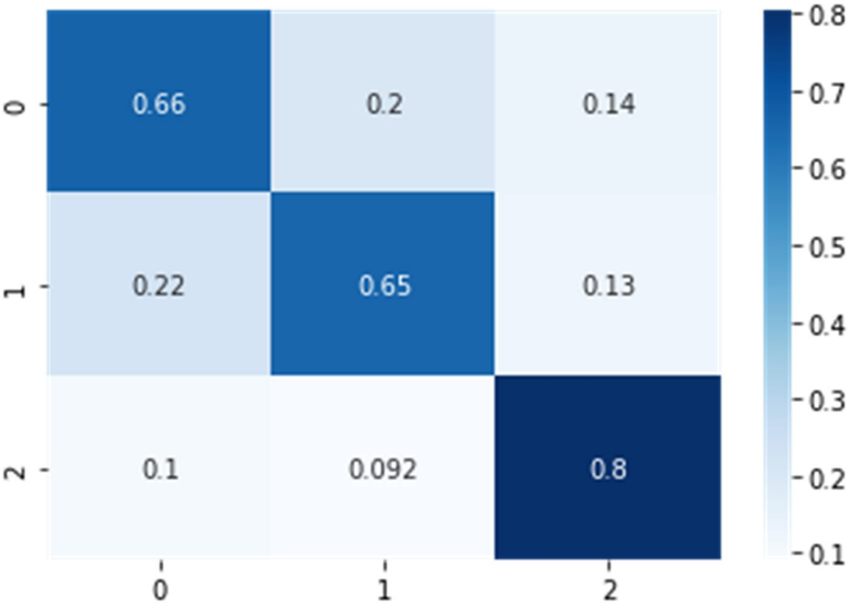
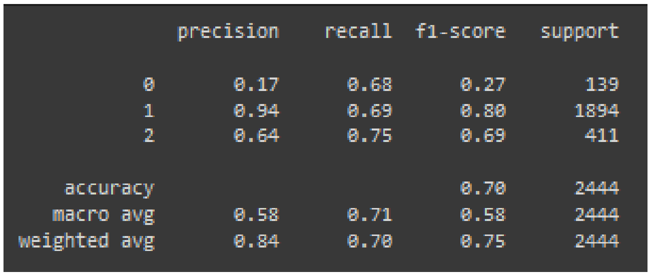


Figure6. Prediction of the context of the tweets (0 means hate, 1 means offensive, 2 means neutral)



```
              precision    recall  f1-score   support

           0       0.17      0.68      0.27       139
           1       0.94      0.69      0.80      1894
           2       0.64      0.75      0.69       411

    accuracy                           0.70      2444
   macro avg       0.58      0.71      0.58      2444
weighted avg       0.84      0.70      0.75      2444
```

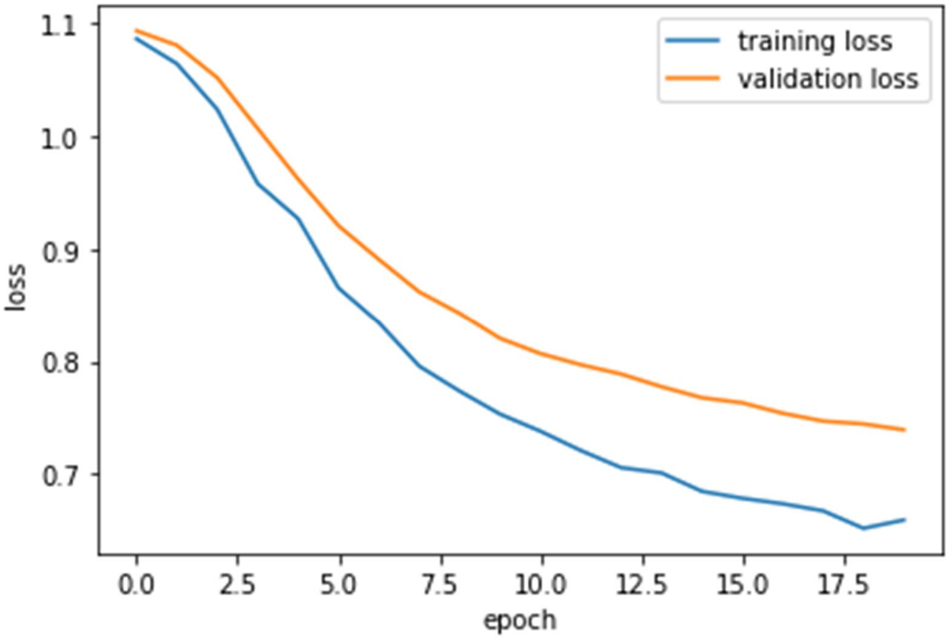


Figure7. Training loss vs validation loss

1.4

As we can see from the results the pretrained hatebert model has better performance on predicting hate tweets. Normal (class 2) and hate tweets (class 0) precision, recall and f1-score are higher in hatebert model because this model is re-trained by specific dataset that mostly contained hate comments.

So, models like Hatebert are generally trained to perform better at classifying of a particular subject for example hate or political comments.

1.5

T5 is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format. T5 works well on a variety of tasks out-of-the-box by prepending a different prefix to the input corresponding to each task, e.g., for translation: *translate English to German:* …, for summarization: *summarize:* ….

T5 uses relative scalar embeddings. Encoder input padding can be done on the left and on the right.

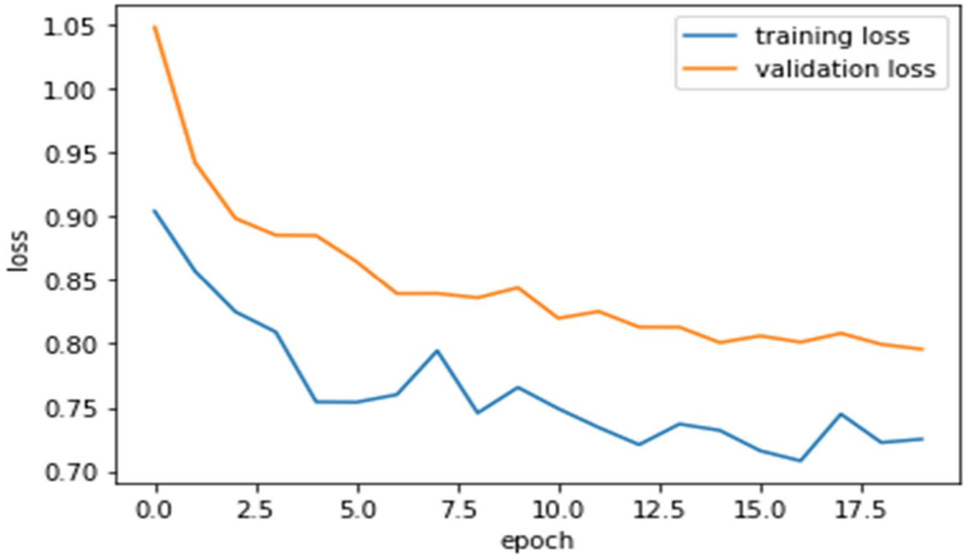


Figure8. Training loss vs validation loss

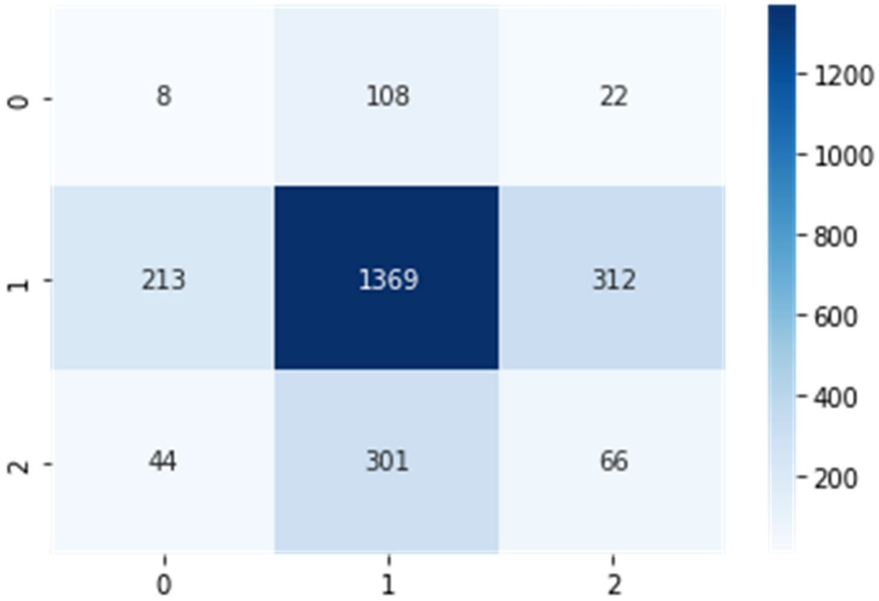
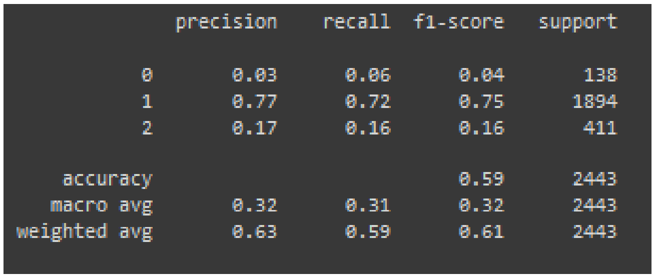


Figure8. Prediction of the context of the tweets (0 means hate, 1 means offensive, 2 means neutral)



```
              precision    recall  f1-score   support

           0       0.03      0.06      0.04       138
           1       0.77      0.72      0.75      1894
           2       0.17      0.16      0.16       411

    accuracy                           0.59      2443
   macro avg       0.32      0.31      0.32      2443
weighted avg       0.63      0.59      0.61      2443
```