Individual Report
Ani Meliksetyan

This project was a big learning experience and opportunity for me.

First, it gave me a chance to work with a group of data scientists and simulate a data science working environment. I started using GitHub which I had never used before, which was again a learning experience.

Second, I went through a number of phases, which at this point serve good reflection points for myself.

In **week 1**, I conducted a lot of research to read about the topic of our choice and how it was treated in academia and industry. I came across some good projects and articles which I shared with the group as sample projects for inspiration. At this stage, they are deleted from our GitHub repository, though.

In **week 2**, I tried to implement a Bi-Directional LSTM architecture following the process described in one of the articles. Because at that time, we had not covered RNNs and LSTM in particular, I was not sure if I had created anything meaningful, although, it would run and print some results.

In **week 3**, I tried to write a classification model using transformer model RoBERTa. At that time, I did not realize that I had developed the model by training and validation sets only, meaning that there would be no data for me to test it. That's said, my code would not run properly, but I shared it with the group, and my code served a basis for further development and optimization which Chris and Timur did. I also worked on using a pre-trained model Pegasus for summarization from Hugging Face. The code I created was not very long and worked pretty well, letting us summarize news articles. It was incorporated into our Streamlit file by Timur. The summarization will be showcased during the presentation.

Additionally, I have been continuously trying to design an LSTM model that would work on our dataset well. In fact, I have created two separate files for LSTM. Originally, I was trying to build the model using custom-made functions for pad_pack_sequence and I was using GloVe embeddings with 50 dimensions. Given that, my model's performance was really poor and I was not sure how to improve that. One option was to use the GloVe embeddings pretrained with higher dimensions, such as 300. Technically, it was very difficult for me to make that code run, and I had to print out every line from the forward method and the training loop outputs to see why I was facing issues with the inconsistency of the shapes. For example, it would take the correct batch size, the sequence length, and the dimensions, however, during the output, the shapes would get distorted, disabling the code to run. Then I was somehow able to run the code, however, it turned out, that some unexpected errors occurred in the predicted labels, going out of range of 0,1,2 and printing numbers such as 262, o4 53, etc. In a nutshell, it was a struggle, and I spent a lot of time trying to figure that out. However, by doing so, I learned a lot about LSTM architecture, the way it runs and classifies.

That's said, I decided to change my custom-made approach and used PyTorch functions for embeddings and for padding and packing. I tried different parameters, such as increasing the length of sequences, vs decreasing the lengths. I also tried to use different learning rates, and varied the number of layers between 2 and 3. That improved my model's performance a bit, however, it is still underperforming transformer models designed by Timur and Chris. Another LSTM model was created by Chris, where he used Word2Vec embeddings and his results seem to be better than my current model. However, I decided to keep this in the project report. Also, since we have decided to use the RoBERTA classifier as our base model, then I worry less about the LSTM underperforming the RoBERTA.

In **week 4**, I did a lot of experimentation, trying different hyper-parametres, trying to use grid search for that, which never worked, and I am not sure if I have to send all that code versions to you. Instead, I have uploaded only the codes that either run properly, or are properly written and ran before being incorporated into the larger models(for example, the Pegasus model for our summarization purposes). I really spent a lot of hours to find the bugs, trying to run them.

I just wanted to thank you for your support during this class, professor Jafari. I do hope I will be able to build on what I have learned and be more efficient in my NLP endeavors.