Progress Report:

Project: Text Classification Competition (Sarcasm on Twitter)

1) Which tasks have been completed?

So far, as part of the project I have completed the following tasks:

1. Take training data into pandas (Python)
2. Processing of the data which includes:
    a. Removal of @USER references
    b. Removal of HMTL references
    c. Removal of hashtag symbols
    d. Removal of links
    e. Removal of punctuation
    f. Removal of stopwords
    g. Removal / Substitution of emojis by text (testing these two scenarios).
    h. Removal of other special symbols.

3. I have created a classifier using Neural Networks, specifically using LSTM with an Embeddings Layer (activation function being sigmoid, which I find appropriate for a classifier).

4. I have tested three approaches / packages:
    a. Word2Vec
    b. Glove
    c. FastText

    So far, I have gotten my best results (69.5% F1 Score) with Fast Text under specific conditions (see below). I was expecting much better results with Glove but that turned out not to be the case.

What am I currently working on?

5. With the NN as described above, I am currently testing different embedding dimensions (25, 50, 100, 200) for the embeddings layer. I have no conclusion yet, but it seems like I am getting better results with Embedding dimension = 100.
6. Currently testing two optimizers:
    a. Started with Adam at learning rate 0.01, results were somewhat overwhelming (F1 score ~55%)
    b. I found better results by using Adamax (which is recommended for NN using embeddings), and at learning rate 0.001. Because of the smaller learning rate, I need to run more epochs (hence is slower to learn, which is totally expected), but I managed to improve to my best result using that.
    c. I am using cross-binary entropy as loss function, which again I find pertinent for the problem I am trying to optimize for (classification issue).
    d. It seems, however, that I am facing a barrier as I cannot move past 70% with my current efforts (see below on what's pending and issues facing).

2) Which tasks are pending?

I have a functioning NN model that is able to assess the tweets as required by the assignment. So in that sense I have something to show 😊.

Having said that, and as mentioned before, I am not achieving the desired result (beating the baseline, currently at above 72.3% F1 Score.

Things I still have pending to do:

- Further research for models that can take into advantage context information (which I am currently not using), such as replies to tweets, which could shed more light on the intention of the original tweet that is subject to sentiment analysis.
- Check other potential approaches to the problem that do not include training a Neural Network (although I am not sure they could be better).

3) Are you facing any challenges?

Yes. I have a working model, but I am not hitting high numbers to beat the baseline, and I would like to do so. Going further I would like to have a high F1 score (likely by having a high precision score, but open to less accuracy for a higher recall, if that gets me above the F1 score). With my current approach I am unable go above 69.5% F1 Score (with various numbers in precision ranging from 55% to 63%).

My other challenges are:

- I am currently not using the context information for tweets given in the training and tests set, and I think I am missing out on important information to improve my algorithms / models. I am not using this information mostly because I do not know how to embed or incorporate that information into the NN to improve the predictions (or how to use it at all even if it were not a NN).
- I am currently converting emojis to text, but I am not sure if this is making any difference (versus, say leaving the emojis untouched and let the NN process them as they are represented).
- I am not sure I am using the best optimizer and loss function (I think I am, but I don't know what else is out there).

So, if you have any guidance, to a paper, specific model in Kaggle, whatever that can be of use for me, please feel free to share that information with me as part of the feedback. I am working in this project alone.