Check in 3 - Dora

1. **Do you have data fully in hand and if not, what blockers are you facing?**

- Yes data is fully in hand

2. **Have you done a full EDA on all of your data?**

- Yes, data has been cleaned
- NaN filled with 'nopost'
- removed posts of inordinate lengths (more than 5k words)
- removed non english posts
- visualised top words and top bigrams for each source (reddit, HWZ, sgtalk)

3. **Have you begun the modeling process? How accurate are your predictions so far?**

- preprocessing done (tokenise, lemmatise, removed stopwords)
  - This only applies for non-BERT models as BERT has its own tokenising method

- tried to label sentiment on data using textblob and vader, gave very inaccurate results
- turned to manual labelling
  - labelled 1000 out of 8000 posts for train set (2000 posts set aside for test set)

- modeling of labelled data (sample of 1000 posts) :
  - classification models (naive bayes, logreg)
  - neural nets: recurrent neural networks
  - BERT

- Accuracy of preds
- based on accuracy score, which is not ideal as classes are unbalanced: 30% negative (class 0), 5% positive (class 2), 65% neutral (class 1) for labelled data.
  - classification models: similar results for NB and logreg
    - train accuracy (.78), validation accuracy (.75)
    - remaining scores below - this is the naive bayes model result, logreg had similar performance

	precision	recall	f1-score	support
0	0.647	0.268	0.379	41
1	0.791	0.973	0.873	148
2	0.000	0.000	0.000	11
accuracy			0.775	200
macro avg	0.479	0.414	0.417	200
weighted avg	0.718	0.775	0.724	200

  - Recurrent Neural Network (yet to do)
  - BERT results
    - Validation loss is very high, which is worrying. But val score seems ok
    - Seems like validation loss increases after first epoch, while training loss decreases, causing high variance. What's the issue?
    - Problem with understanding how to finetune BERT to achieve a better score.

epoch	Training Loss	Valid. Loss	Valid. Accur.	Training Time	Validation Time
1	0.73	0.66	0.75	0:01:19	0:00:07
2	0.61	0.64	0.75	0:01:19	0:00:07
3	0.45	0.64	0.75	0:01:19	0:00:07
4	0.30	0.78	0.76	0:01:19	0:00:07
5	0.21	0.80	0.78	0:01:19	0:00:07
6	0.16	0.84	0.75	0:01:19	0:00:07
7	0.13	0.94	0.77	0:01:19	0:00:07
8	0.08	0.97	0.75	0:01:19	0:00:07
9	0.05	1.02	0.77	0:01:19	0:00:07
10	0.04	1.02	0.76	0:01:19	0:00:07

Check in 3 - Dora

Comment:
I am thinking of not using BERT and just using LSTM if its easier to manage and optimise. There are alot of intricacies with BERT that I'm do not understand, and I feel like I may not have the time to understand it due to the time constraints of the capstone project.

Suggestion:
Drop BERT if LSTM model gives similar results?

4. **What blockers are you facing, including processing power, data acquisition, modeling difficulties, data cleaning, etc.? How can we help you overcome those challenges?**

- Problem with getting optimal result in BERT
- Issue with relatively low validation score after modelling sample data
  - is it ok to continue if val score hovers at 0.75?

5. **Have you changed topics since your lightning talk? Since you submitted your Problem Statement and EDA? If so, do you have the necessary data in hand (and the requisite EDA completed) to continue moving forward?**

- No changes

6. **What is your timeline for the next week and a half? What do you *have* to get done versus what would you *like* to get done?**

Next steps for sentiment analysis are to:
1. Decide on a model
2. Use model to label remaining train data (80% of all data)
3. Check accuracy of predictions
4. Fit model again, optimise params
5. Predict sentiment of posts in test data
6. Check accuracy of test data predictions
7. Fit production model.

Aim to finish by mid-Week 10

Thereafter I can move on to:
- topic modelling (by end of week 10, early-week 11)
- deployment model (end-week 11)
- cleanup, readme, presentation slides (week 12)

8. **What topics do you want to discuss during your 1:1?**

- discuss BERT model performance
- how to finetune BERT
- using google collab GPU - when to use it?
- Validation score of .75 acceptable?