**Meeting Minutes**

**Initial meeting: 21/10/19:**

* How to articulate Initial work plan and problem description was discussed. Main data for fine-tuning the hate speech detector (from the 2019 hateval competition on codebase) was difficult to source – as I wasn’t yet accepted as a participant in the HatEval competition. So, Barry provided his copy of said data
* Research on transfer learning was advised. This is the method by which BERT transfers it’s already gained knowledge on natural language to a downstream task such as classifying hate speech
* Research on Fine-Tuning also advised. This is the method by which BERT is more specifically tailored towards binarily classifying hate soeech
* Discussed was the difficulties in ascertaining hate speech data. This could be an interesting point in the initial project description. At the very least it shows meticulous data inspection which is always good to see
* Explore methods used by related studies, whether it is methods of test pre-processing or machine learning frameworks. Especially on the approaches explored in the HatEval competition which will suit as an appropriate benchmark to compare your performance against.
* Do deep research into the BERT paper so that you know what framework you’re dealing with and how to best utilise it.

**Meeting 22/11/19**

* Chance meeting in queen’s computer science lab. Discussed was the inability of my PC to Run BERT as the processing power was too small. Computer science lab GPUs, Google Colab TPUs and codalab website were a suggestion.
* Using the BERT-Base is an option, as well as reducing batch size

**Meeting 2/12/19**

* BERT-Base has been configured to work upon GPUs in Google Colab and have successfully classified hate speech datasets.
* Namely the hate speech dataset from Analyticsvidhya.com, 20th place was attained in the contest from around 1,000 entries upon little to no parameter tuning and very basic text pre-processing
* Text pre-processing methods were discussed including emoji replacement and hashtag segmentation

**Meeting 7/12/19**

* Layout and approach to initial project Demo discussed
* As a way to further improve upon BERT word embeddings, further pre-training BERT upon a massive number of tweets using masked language modelling, much like BERT was pre-trained may yield significant performance increases as BERT could become better adapted to understanding the language used in tweets.
* Further pre-training should not be identical to the way BERT was pre-trained, next sentence prediction should not be used as it is not relevant to classifying individual tweets. If there is a way to turn off this task in the further pre-training of BERT that would be advised

**Initial Demo 12/12/19**

* Project code and findings thus far were demonstrated in a powerpoint presentation

**Meeting 20/1/20**

* Further pre-training data shown. This stage is ready to run, just going over some finer details on what learning rate in further pre-training sounds appropriate
* Emoji replacement and hashtag segmentation showcased. These methods seem to harm performance however
* Also discussed is the upcoming HASOC competition. Could be a good opportunity to benchmark my classifier against other participant’s systems.
* System is now in the top 10 of AnalyticsVidhya.com hackathon and just shy of the best classifier among those who have attempted to classify the HatEval dataset.

**Meeting 14/2/20**

* Non-determinism of Tensorflow means that results collected thus far are not reliable. The variance from run-to-run is greater than the performance increase changes bring. Dr Devereux has suggested a novel way to reduce this variance by using a form of nested cross-validation. Cross validating a sample of the data 5 times will reduce variance and give more of a reliable metric that I can use to judge changes in parameters or pre-processing pipelines.
* Also, grid search is not an appropriate method to finding the best hyper parameters for a neural network, especially one as large as BERT. Instead an early stopping procedure should be implemented wherein a model is trained for a certain amount o steps then checkpointed every n steps and evaluated against a holdout development set using a metric of your choice.
* Tweet datasets are sourced via in-common hashtags such as #Charlottesville or #immigrationorder which obviously don’t appear often in the general corpora of tweets relative to how often they appear in these tweet datasets. My concern was that the word masking task in further pre-training may learn to over predict these terms and therefore learning of natural language would be harmed. Dr Devereux understood this concern but expressed reservations on this technique because although these terms may be overpopulated in the dataset, their removal might result in a lot of harm in terms of context of the tweets and their semantic meaning. This may be more deleterious to natural language understanding. An analysis of whether to remove these terms from further pre-training or not was suggested

**Meeting 27/2/20**

* Meeting in ECIT building as per Dr Devereux’s recommendation. Here I was able to meet with PhD and masters students who have a great amount of expertise in NLP. They recommended another form of emoji translation where I should translate emojis into unique tokens which then can be hard coded into BERT’s vocab file. Also encoded should be token representations of multiple consecutive emojis, which appear often in tweets.
* Novel cross-validation method was met with approval. Advice is that adding a couple of layers to process BERT word embeddings may be beneficial

**Meeting 11/3/20**

* Advice was asked on how to implement neural network architectures in Tensorflow – how to appropriately define layers, activation function, dropout etc. Also if bi-LTSM is a good idea to implement.
* Results were showcased - reliably attained from novel CV method. Hashtag segmentation and emoji replacement improve performance. Replacing emojis with words seems to be preferable to replacing with unique tokens, an initial learning rate of 5e-5 is marginally better than 2e-5 and not removing in-common terms (hashtags) is better than removing them for further pre-training data.
* Different loss functions to be attempted was discussed. They might make an impact, albeit a small one. Would be useful in the AnalyticsVidhya datset which is heavily imbalanced.

**Meeting 30/3/20**

* Zoom meeting – could not be in person due to quarantine
* 23 slide powerpoint detailing collection of results and findings so far
* Organisation of project was stressed. Need to be less results focussed and more orientated towards explaining why I made decisions on strategies employed in each stage. Try to back up your thinking on available research. Don’t just attempt techniques because you can, do it if it hasn’t been done in research before or if you think you can improve upon previous methods
* Clear interpretation and demonstrating deep understanding of your classifier is more important than results
* Dr Devereux thinks my findings might be eligible to be published in an academic journal. He suggests that I should attempt this.