Project Update

Daniel Attia – attia.22
Ryan Baxter – baxter.243
Jonathan Chi – chi.171
Devan Mallory – Mallory.115
Yiming Liu – liu.8672

The project update will consist of your progress towards the final project.

Below is a suggested format for this update:

Analysis about the data: our dataset is English text from Twitter data. Each line of data contains an indicator of whether the text is irony (1) or not (0). The train data is categorized into 3 sections: one containing just the text, one containing the text and emojis, and the last containing the text, emojis, and hashtags such as #irony and #not. The test data is divided into similar categories minus the one containing hashtags as it would invalidate the testing.

How large is your dataset: Our training dataset contains 3834 lines of sentences, and the test dataset contains 784 lines.

How you have plan to evaluated (train/test split ratio, cross-validation)
The data is already split into train and test data, with a ratio of 5 to 1 train to test.

Have you implemented any algorithms yet?
No.
If not, what is you planned algorithm:
We will first import the data from the training text, and we'll be using Naive Bayes algorithm for tweet classification, similar to homework 2, then use the text data set to check accuracy. Because the data contains the distinction of irony or not, we must split the train data into an irony set and a non-irony set for training. We have also considered using the TensorFlow library and implementing a neural network to classify the data (Is this worth considering? Or rather is it within the scope of this project?).

However, this report is open ended (just like the proposal) and you can present your current progress in your own proffered way

You just need to submit a pdf file with your progress, you do not need to submit any code now. You will submit the code for this at the very end.

Only one person of the full group needs to submit this report.

Project proposal: https://competitions.codalab.org/competitions/17468
Dataset: https://github.com/Cyvhee/SemEval2018-Task3/tree/master/datasets