PROJECT PROGRESS REPORT

Sandeep Nanjegowda – sgn3 (**Captain**)

Sunitha Vijayanarayan - sunitha3

Valentina Mondal – vmondal2

1. Progress made so far

Have tried the below models so far:

- Linear Regression
- Naïve Bayes
- LSTM (long-short term memory)
- GRU (Gated recurring units)
- CNN (Convolutional Neural Network)
- SVM (Support Vector Machine)

Have tried word embeddings like glove and word2vector on the models: LSTM, GRU and CNN.

Both test and training set data was initially processed to remove html tags, punctuations, numbers, single characters, multiple spaces. Models were trained and fitted using the given twitter trained set data by splitting it into training and test set. Further the model was applied on the test data set to find out the precision, recall and F-Score.

Linear Regression model performed well on the test set data and gave precision of 0.658, recall of 0.808 and F-score of 0.725. Were able to beat the baseline using the LR model. However, on the other models such as LSTM, GRU, CNN after trying out with both the word embedding – glove and word2vec, got almost the same results and not able to beat the baseline. Got F-score up till 0.69 using the above-mentioned neural network methods.

For SVM, after using cross validation and grid search to find the good hyperparameter for SVM model, still the model could not beat the baseline with the test set data. Got a F-score of around 0.68.

2. Remaining tasks

- Looking to explore BERT (Bidirectional Encoder Representations from Transformers) which is a neural network-based technique for natural language processing pre-training.
- We have extracted additional features such as word counts, number of has tags, number of emojis in each of the tweet and used them in Logistic Regression. We are planning to use these features in Neural Network Models.
- Project documentation and presentation.

3. Any challenges/issues being faced

Unable to beat the baseline (i.e., F-score of 0.723) in live lab leaderboard using LSTM, GRU and CNN.