Progress Report

By: Pinky Chauhan

Topic: Fake news classification using machine learning

1) Which tasks have been completed?
   - As per the recommendation of project proposal reviewer in CMT, I changed the dataset to the one suggested by the reviewer https://www.kaggle.com/c/fake-news/data
   - I have been acquainting myself with different classification algorithm details and also nltk, sklearn, pandas libraries to work on this project.
   - Data Analysis is complete using matplotlib, nltk sentiment analyzer and manual run-through to understand the observations listed in the dataset and its contribution towards the classification.
   - Preprocessing of data is done using nltk to setup training and test datasets, handle missing values, performing tokenization, removing stop words, lemmatization, encode categorical variables as needed.
   - Feature Selection to keep only the most relevant variables that are used for training.
   - Vectorization using sklearn libraries to map words to a corresponding vector of real numbers to find word similarities, etc.
   - Model design and training using several classification algorithms using sklearn libraries (Naive-Bayes, Decision tree and Logistic Regression so far)
   - Data and preliminary notebook are available in Github repo.

2) Which tasks are pending?
   - Models hyperparameter tuning and validation to assess the accuracy and avoid overfitting.
   - Performance evaluation of the different model algorithms used: compute and analyze the metrics precision, recall, F1 score, etc.
   - Create API/script that will take news text as input and generate its classification as real or fake as the result.
   - If time permits, will also try to add a submission of this notebook on Kaggle and evaluate accuracy against other submissions.

Task	Status
Understand classification algorithms in depth and familiarize with nltk (I am new to machine learning world and will need to research/obtain a deeper understanding)	Complete
Environment setup	Complete
Data analysis and preprocessing	Complete
Feature selection and vectorization	Complete
Model design, training and hyperparameters tuning	Model design and training complete; Tuning in progress
Testing and evaluation	In progress

Integration with final output script/API	To be done
Prepare presentation	To be done

3) Are you facing any challenges?
   - Nothing major at this time. I am relatively new to machine learning, NLTK, sklearn libraries. But there is good information available online and that has been very helpful thus far.