Capstone Project Proposal

1 Domain Background

This project relies on a Kaggle competition "Real or Not? NLP with Disaster Tweets" that its goal is to classify the tweets whether it is real or fake disasters. It would be very beneficial to solve this problem as it would help the relief organizations or news agencies to know about these disasters and try to help the people or report the news. And I find it very useful for me as my work might help others and it will enrich my knowledge by learning more about NLP.

2 Problem Statement

The problem is to classify the tweets whether it a real or fake tweets about disasters. The proposed solution is to use classifiers like XGB, SVM or even NN.

3 Datasets and Inputs

The dataset is provided by Kaggle.com. It is divided into two sets.
1)train.csv
Its shape is (7613, 5)
The features are:
1- id
2-text: the text of the tweet.
3-location: the location where the tweet was sent from.
4-keyword: a particular keyword from the tweet.
5-target: this denotes whether a tweet is about a real disaster (1) or not (0)

2)test.csv
Its shape is (3263, 4)
It has the same features but without the target feature.

So the total number of the data instances including the train.csv and test.csv is 10876 instance
I would use 20% of train.csv for validation.

4 Solution Statement

I would go through the steps of solving a machine learning problem starting from EDA till choosing the best model from (NN, XGB, SVM).

5 Benchmark Model

I would choose Linear Learner as the benchmark model

6 Evaluation metrics

I choose between ROC/AUC and precision/recall metrics to measure how the model is doing.

7 Project Design

The workflow of the project depends on two resources:
1)The notebooks of previous Udacity projects
2)"Hands on Machine Learning with Scikit learn and Keras and TensorFlow" textbook
The steps would be :

1)Data Cleaning
* Drop the columns that would make the model deduce wrong information like the id.
* Compute the medium for columns and fill the null data if it numeric.
* Fill the null instances of data with appropriate values .
* I would use Corr matrix to have more insights about the data and see the more correlated features to the the wanted target.

2)Use NLP techniques to make the data more appropriate for ML algorithms
*Refine the tweets from any insensible symbols like html tags or URLs.
*Convert the words to lowercase.
*Use PorterStemmer from nltk module to get the stem of the words.
*Use word embedding like word2vec or Glove to convert the words into more sensible vectors for the algorithm.

3)Select fixed size for the batches.
*If the size of the tweet is bigger than the selected size. It would be cut, Otherwise it would be filled with non used values to indicate that is just a filler.
*It is best to choose fixed size for the batch in case of using NN.

3)Train the ML model. The considered algorithms would be NN, XGB, SVM.