Subject: Proposal towards Udacity Machine Learning Advanced Nanodegree Program Capstone Project.

Author: Prashant Tripathi (xprashanttr@gmail.com)

Introduction:

I have chosen "Toxic Comment Classification Challenge" on Kaggle for my Advanced Machine Learning Nanodegree Capstone Project.

Kaggle Competition title: Jigsaw Toxic Comment Classification Challenge

Kaggle Competition URL: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

Domain Background :

The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

The Conversation AI team, a research initiative founded by Jigsaw and Google (both a part of Alphabet) are working on tools to help improve online conversation. One area of focus is the study of negative online behaviours, like toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). So far they've built a range of publicly available models served through the Perspective API, including toxicity. But the current models still make errors, and they don't allow users to select which types of toxicity they're interested in finding (e.g. some platforms may be fine with profanity, but not with other types of toxic content).

My Motivation: I am somehow inclined towards text analytics and processing. In 2015, I did a basic reading on text analytics and came across many concepts on why text universe is different from other areas of analytics. I am a database professional and have faced many problem statements and feature requirements in my career, where appropriate solution would have been 'Context' search but because of technical limitations of tools and platforms, we went with simple string search.

In medium term – I want to build a classifier for fake news detection. This project is starting step.

Problem Statement :

To build a multi-headed model that's capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate. This is a multi-label classification problem whereby each comment can belong to one or more class. There are 6 classes as given below.

toxic	severe_toxic	obscene	threat	insult	identity_hate

Appropriate prediction of classes for given comments can have multiple benefits, biggest of them would be creating social media platforms more censored hence suitable for everyone in today's connected world.

Datasets and Inputs :

Dataset provided is of comments from Wikipedia's talk page edits.

Disclaimer: the dataset for this competition contains text that may be considered profane, vulgar, or offensive.

Data provided has large number of Wikipedia comments which have been labelled by human raters for toxic behaviour. The types of toxicity are:

toxic	severe_toxic	obscene	threat	insult	identity_hate

File descriptions:

train.csv - the training set, contains comments with their binary labels. Below is structure.

id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate

test.csv - the test set, model must predict the toxicity probabilities for these comments. To deter hand labelling, the test set contains some comments which are not included in scoring. Below is structure.

id	comment_text

sample_submission.csv - A sample submission file in the correct format. Below is structure.

id	toxic	severe_toxic	obscene	threat	insult	identity_hate

Solution Statement :

I intend to solve this problem statement in below steps (high level steps):

(a) Embedding of given comments.
(b) Model each type of toxicity so that other comments can be assigned a class probability.

Training the model and Predicting the class probabilities are the next steps.

I intend to use tfidf vectorizer for word embedding, and a 2 stage classifier – RF and NN in the same flow for training the model. I intend to use stratified shuffle split to divide provided training data(train.csv) into model-training and validation sets.

Multiple runs will be required for parameter tuning. I intent to run this model on top of test data(test.csv) provided in competition and evaluate my submission on Kaggle also.

**Assumption: The underlying assumption is that training data set has been accurately classified, and has enough coverage to build a generic model.*

***Implementation is subjected to change depending on real time situations faced while developing the solution.*

Evaluation Metrics :

I intend to use two metrics – MSE and ROC AUC score to evaluate this model, since this is a multi-label classification problem.

I intend to achieve AUC of min 0.80 and MSE of max 0.25.

Benchmark Model :

As part of Conversation AI project (a collaborative research effort exploring ML as a tool for better discussions online) , There are existing text classification models , existing datasets etc. One of the project I came across was on Github : https://github.com/conversationai/unintended-ml-bias-analysis

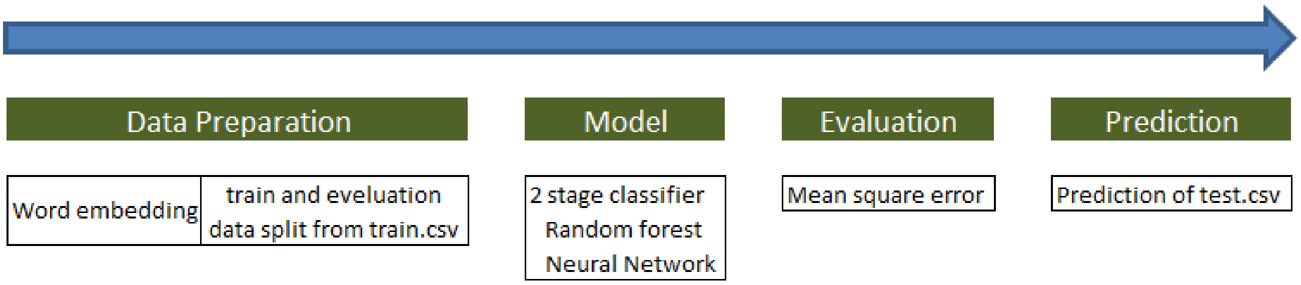
This reference is mentioned in Kaggle Competition Description page.

To tag any model as benchmark seems little difficult to me, as models mentioned in the link above neither have mention of any training data nor any quantitative attribute to evaluate any code following any approaches. Any suggestions are welcome here.

However, in my solution, I intent to achieve AUC of min 0.80 and MSE of max 0.25 as mentioned above.

Project Design:

Below is a high level flow diagram of solution.



Data Preparation

Word embedding

train and eveluation data split from train.csv

Model

2 stage classifier
Random forest
Neural Network

Evaluation

Mean square error

Prediction

Prediction of test.csv

Also, below is count of some scenarios to give an idea of how data is keeping in training set provided.

High level data counts	
Total number of records	159571
Toxic	15294
severe_toxic	1595
Not classified as Toxic but classified as severe_toxic	0
obscene	8449
threat	478
insult	7877
identity_hate	1405
classified in one or more classes	16225
not classified in any class	143346