Intro to Machine Learning Project Proposal

Project Title: Quora Insincere questions classification

Project synopsis:

Quora is generally used for the purpose of knowledge sharing, where people across continents can put forth any questions that they have, and experts/ non-experts in that domain can revert back with an answer.

However, there are some users who through their question intend to make a statement rather than looking for helpful answers, or post questions which are completely absurd or have no factual relevance.

These questions can also become a threat to the concept of a safe internet as some of these questions can imply negative opinion or promote hatred towards certain sections of society. Quora terms such questions as Insincere, and lately hosted a [kaggle competition](#) to identify such questions on their platform.

In this project, we plan to build a machine learning classifier to help classify insincere questions from the dataset shared by Quora on kaggle.

Notes on the dataset:
  1. The dataset is skewed as the number of sincere (or relevant) questions far outnumber the number of insincere (or bad) questions.
  2. The annotated labels also do have noise, which means certain data samples might have an incorrect label associated to them.

Sample sincere questions:
  What is the best place to buy maths/physics T-shirts?
  Which book is good for power electronics?
Sample insincere questions:
  How do I break your foot?
  Is it mostly people who have no life who use Quora?

Planned Experiments:
We plan to employ both traditional machine learning and deep learning based experiments to build our target classifier.

Feature Extraction:
  1. We intend to use **tf-idf (Term frequency - inverse document frequency)** scores as feature extraction module for the given dataset when experimenting with conventional machine learning algorithms and single layer neural nets.

2. While, we plan to use **Word vector embeddings** for training convolutional and recurrent neural nets on the problem.

Classifiers to experiment with:

We plan to do a progressive study, where we begin from Machine learning models with the least number of parameters such as logistic regression and progress towards models with more expressive ability such as SVMs and Neural nets. Summarily, we plan to follow the following roadmap:

1. Logistic regression with ridge and/or lasso regularization on tf-idf scores
   a. We additionally plan to survey here different loss functions such as hinge and cross entropy loss, for their effect on the performance of logistic regression.
2. SVM on tf-idf scores:
3. Single layer neural net with ridge and/or lasso regularization on tf-idf scores
   a. Similar to logistic regression, we plan to survey different loss functions such as hinge and cross entropy loss, for their effect on convergence and performance of the classifier.
4. CNN on word vector embeddings
   a. CNNs have been shown to perform incredibly well on natural language processing tasks, such as sentence classification, due to their prowess of capturing local features [1]. Hence CNN might be a promising classifier for the same.
5. RNN (a net of LSTM layers) on word vector embeddings: RNNs due to recurrent connections between RNN cells have profound capability to capture temporal relationship [3]. And since a question or any natural language sentence is a sequence of words, hence RNN architecture is a must for experimenting to solve this problem.

References:

1. Quora Insincere questions classification:
   https://www.kaggle.com/c/quora-insincere-questions-classification/
2. Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).
3. The Unreasonable Effectiveness of Recurrent Neural Networks -- Blog by Andrej Karpathy http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Team Members:

Aniket Bhatnagar (ab8700), Dhruv Goyal (dg3278), Yudhishthir Deshpande (yd1568)

Member responsible for uploading submission:

Aniket Bhatnagar (ab8700)