Machine Learning Engineer Nanodegree

Capstone Proposal: Quora Insincere Questions Classifier

Karim Fateem

January 17th, 2019

Proposal

Domain Background

A major challenge faced by social media, news and content websites is maintaining constructive non-toxic dialogue between users. The notion of what is considered constructive or non-toxic itself is a subjective determination and often varies across cultures. Social media companies spend significant resources on developing and enforcing community standards and are increasingly turning to machine learning (ML) to assist in applying those community standards.

One such company is Quora, a platform that empowers people to learn from one another. On Quora, people can ask questions and connect with others who contribute unique insights and quality answers. Quora has increasingly relied on ML in various parts of its products. Quora recently launched a Kaggle competition to classify insincere questions, which along with my past efforts at Parlio towards building a platform for civil discourse, are the motivations for this capstone project.

Problem Statement

This project aims to classify whether a question asked on Quora is sincere or insincere. An insincere question is defined as a question intended to make a statement rather than to solicit helpful answers. Characteristics used for determining if a question is insincere are discussed in the “Datasets and Inputs” section below. The dataset includes labeled items. The solution model’s accuracy and F1-score will be measured against the validation dataset.

Acceptable sentiment analysis accuracy scores fall between 0.7 and 0.9 according to industry practitioners and the F1-score can be compared against leading submissions on Kaggle leaderboard.

Datasets and Inputs

Quora provides training data that includes the question that was asked, and whether the question was identified as insincere (target = 1). The ground-truth labels contain some amount of noise, since there is subjectivity when labelling a question as sincere or insincere. Some characteristics used to determine insincerity are listed below.

- Has a non-neutral tone
  - Has an exaggerated tone to underscore a point about a group of people
  - Is rhetorical and meant to imply a statement about a group of people
- Is disparaging or inflammatory
  - Suggests a discriminatory idea against a protected class of people, or seeks confirmation of a stereotype
  - Makes disparaging attacks/insults against a specific person or group of people
  - Based on an outlandish premise about a group of people
  - Disparages against a characteristic that is not fixable and not measurable
- Is not grounded in reality
  - Based on false information, or contains absurd assumptions
- Uses sexual content (incest, bestiality, pedophilia) for shock value, and not to seek genuine answers

The Quora Kaggle challenge also permits the use of the following word embeddings:

- GoogleNews-vectors-negative300 - https://code.google.com/archive/p/word2vec/
- glove.840B.300d - https://nlp.stanford.edu/projects/glove/
- paragram_300_sl999 - https://cogcomp.org/page/resource_view/106
- wiki-news-300d-1M - https://fasttext.cc/docs/en/english-vectors.html

The dataset can be downloaded from the following Kaggle competition link: https://www.kaggle.com/c/10737/download-all

Solution Statement

The classifier will be implemented using a Recurrent Neural Network (RNN), possibly a Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) network with attention mechanisms and a word embeddings layer.

Benchmark Model

The following CNN with a GloVe embedding will be used as a benchmark: https://www.kaggle.com/yekenot/2dcnn-textclassifier

The model has a F1-score of 0.671 and accuracy > 0.95.
The leading F1-score in the Kaggle leaderboard is 0.711 as of the time of writing this proposal.

Evaluation Metrics

Given:

- *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.
- Precision = $\frac{TP}{TP + FP}$
- Recall = $\frac{TP}{TP + FN}$

The benchmark and solution models will be measured using the following metrics:

- Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$
- F1-score = $2 \times \frac{Precision \times Recall}{Precision + Recall}$

Project Design

Analysis and preprocessing:

One or more of the permitted word embeddings will likely be used in the embedding layer of the network. In this preprocessing phase, embeddings:

- will be analyzed to determine the Out Of Vocabulary (OOV) rate, and
- mutated to reduce OOV.

Additionally, the input dataset will be sampled to determine whether some particular data augmentation techniques could be used.

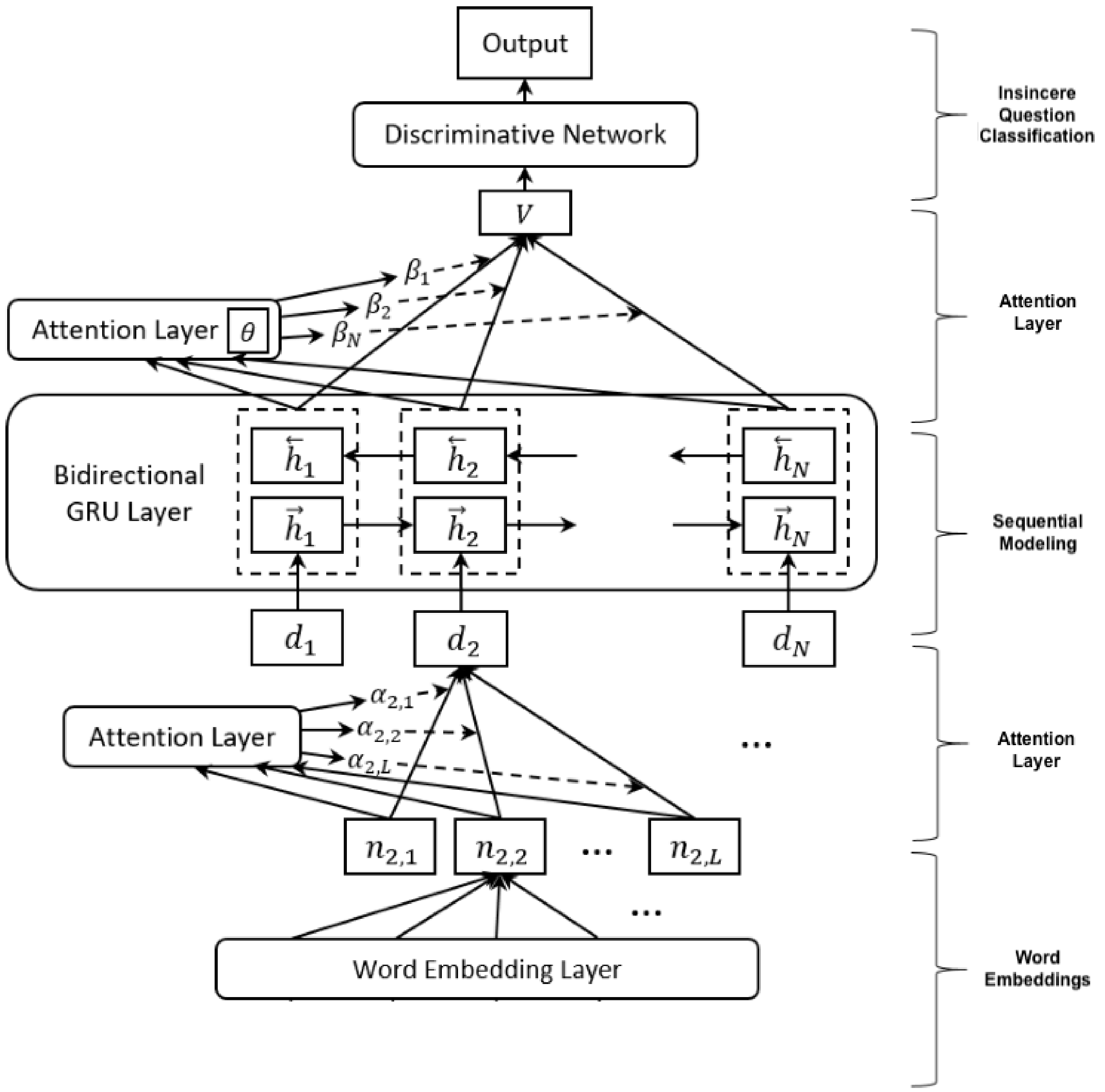
Embedding layer:

In the most naive implementation, one of the embeddings will be selected as input to the network's embedding layer. A more complex solution may attempt to combine multiple embeddings using a CNN. The output of the CNN could be used as the input to the embedding layer.

Building RNN model:

- Training will be conducted using a LSTM and GRU in an attempt to select the best performing network.
- Regularization techniques, such using regularizers and dropouts, will be used to reduce overfitting. Data augmentation may be explored to prevent overfitting.
- Attention layers may be introduced to reduce the "Vanishing Gradient Problem."

The final neural network could be illustrated as follows:


Output
Discriminative Network
V
β_1
β_2
β_N
Attention Layer
θ
Bidirectional GRU Layer
$\overleftarrow{h}_1$
$\overleftarrow{h}_2$
$\overleftarrow{h}_N$
$\overrightarrow{h}_1$
$\overrightarrow{h}_2$
$\overrightarrow{h}_N$
d_1
d_2
d_N
$\alpha_{2,1}$
$\alpha_{2,2}$
$\alpha_{2,L}$
Attention Layer
...
$n_{2,1}$
$n_{2,2}$
...
$n_{2,L}$
...
Word Embedding Layer
Insincere Question Classification
Attention Layer
Sequential Modeling
Attention Layer
Word Embeddings

Works Cited

Quora Insincere Question Challenge | Kaggle, www.kaggle.com/c/quora-insincere-questions-classification.

Improve Your Score with Some Text Preprocessing | Kaggle, www.kaggle.com/theoviel/improve-your-score-with-some-text-preprocessing.

10153181162182282. “Illustrated Guide to LSTM's and GRU's: A Step by Step Explanation.” *Towards Data Science*, Towards Data Science, 24 Sept. 2018, towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21.

Hu, et al. “Listening to Chaotic Whispers: A Deep Learning Framework for News-Oriented Stock Trend Prediction.” *[Astro-Ph/0005112] A Determination of the Hubble Constant from Cepheid Distances and a Model of the Local Peculiar Velocity Field*, American Physical Society, 16 Mar. 2018, arxiv.org/abs/1712.02136.

Nabi, Javaid, and Javaid Nabi. “Machine Learning - Word Embedding &amp; Sentiment Classification Using Keras.” *Towards Data Science*, Towards Data Science, 4 Oct. 2018, towardsdatascience.com/machine-learning-word-embedding-sentiment-classification-using-keras-b83c28087456.

“Parlio.” *Wikipedia*, Wikimedia Foundation, 16 Dec. 2018, en.wikipedia.org/wiki/Parlio.

“Quora - A Place to Share Knowledge and Better Understand the World.” - *Quora*, www.quora.com/.

Sebastian Ruder. “Word Embeddings in 2017: Trends and Future Directions.” *Sebastian Ruder*, Sebastian Ruder, 24 Oct. 2018, ruder.io/word-embeddings-2017/index.html#oovhandling.

Synced. “A Brief Overview of Attention Mechanism – SyncedReview – Medium.” *Medium.com*, Medium, 25 Sept. 2017, medium.com/syncedreview/a-brief-overview-of-attention-mechanism-13c578ba9129.

Tien, et al. “Sentence Modeling via Multiple Word Embeddings and Multi-Level Comparison for Semantic Textual Similarity.” *[Astro-Ph/0005112] A Determination of the Hubble Constant from Cepheid Distances and a Model of the Local Peculiar Velocity Field*, American Physical Society, 21 May 2018, arxiv.org/abs/1805.07882.

Yang, Zichao, et al. “Hierarchical Attention Networks for Document Classification.” *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, doi:10.18653/v1/n16-1174.

Zhang, Ye, et al. “MGNC-CNN: A Simple Approach to Exploiting Multiple Word Embeddings for Sentence Classification.” *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, doi:10.18653/v1/n16-1178.

Zhao, Chuanjun, et al. “Deep Transfer Learning for Social Media Cross-Domain Sentiment Classification.” *Communications in Computer and Information Science Social Media Processing*, 2017, pp. 232–243., doi:10.1007/978-981-10-6805-8_19.