Quora Insincere Questions

Submitted by Ankit Maurya (2018IMT-018)
Under the mentorship of Prof Anuraj Singh

Outline

1. Introduction
2. Literature Review
3. Objective
4. Schematic Flow of the Project
5. Data Prepossessing
6. Implementation
7. Data Modeling: Recurrent Neural Network/LSTM
8. Model Evaluation
9. Future Work
10. Conclusion
11. References

Introduction

Quora is a common online community where people can ask questions and (usually) receive thoughtful responses.

While the majority of questions are posed in good faith, a small number of bad actors ask insincere or problematic questions (for example, questions founded upon false premises, or ones that just intend to make a statement of some kind).

In order to improve the overall group experience, a challenge is to create a classifier that will take a user-generated query as input and automatically identify it as sincere or insincere.

Literature Review

In their paper "Exploring Deep Learning in Semantic Query Matching," Ashwin Dhakal and his co-authors used an Artificial Neural Network approach to predict semantic coincidence between question pairs, extracting highly dominant features, and evaluating the likelihood of a duplicate question on Quora. Words and phrases are mapped into real-number vectors in their research, then feature engineering is applied, which involves NLTK mathematics, fuzzy wuzzy features, and Word mover distances combined with vector distances. Hence, the research has discussed and following the architecture used by the Quora itself along with the knowledge of natural language processing and machine learning.

Literature Review

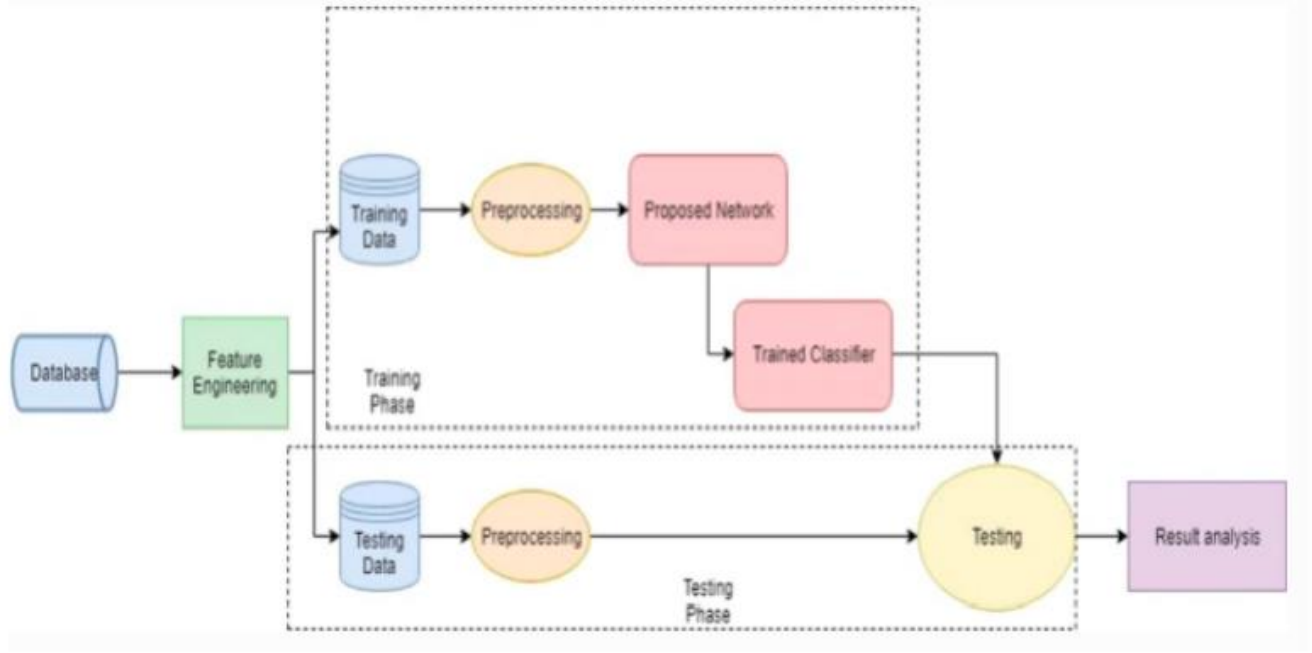
According to a CMU (Carnegie Mellon University) scholar, the paper titled "Hierarchical Attention Networks for Documents Classification" was co-authored by CMU and Microsoft Research in 2016. Six large-scale Datasets were used in the experiment. The tests revealed that this model outperforms the previous best baseline approaches by 3.1 percent and 4.1 percent for smaller datasets like Yelp 2013 and IMDB, respectively. On Yelp 2014, Yelp 2015, Yahoo Answers, and Amazon Reviews, this model outperforms the previous best model by 3.2 percent, 3.4 percent, 4.6 percent, and 6.1 percent for large datasets.

Objective

The main goal of this project is to determine if a Quora query is sincere or not. Insincere questions aren't intended to elicit useful information on a subject of interest; instead, they may be rhetorical or claims. People can tell them apart from normal, sincere questions because of this peculiarity:

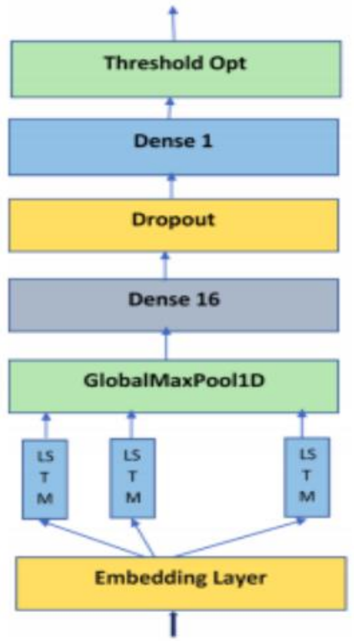
- The tone, which is exaggerated to attract attention to a point or a position rather than being non-neutral.
- Their ultimate intent is to inflame, suggesting a discriminatory idea, seeking confirmation of a stereotype, insulting a person or a group, often basing upon characteristics that are not measurable
- They are generally not based on evidence, nor have a solid connection with reality
- They can contain sexual content in order to be more offensive and elicit a wider response (disdain or shock) from the user community.

Schematic Flow of the Project



Schematic Flow of the Project

1. Model consists of word embeddings, followed by Bi-Directional LSTM layer, and a few dense layers to make final prediction.
2. Global Max Pooling layer helps to detect match at any timestep instead of taking results from the last node.
3. To minimize overfitting, I used a dense layer followed by a dropout layer.



Data Prepossessing

1. The first step will be to analyze and clean the data..

   Html tag removal

   Punctuation removal

   Tokenization

   Lemmatization

   Contraction mapping
2. We divided the training dataset into train and val samples for analysis and data cleaning.
3. We will fill up the missing values in the text column with 'na'.
4. After that make word in lowercase, remove numbers, remove punctuation, remove stop words, pad the sentence and tokenize the text column and convert them to vector sequences.
5. On certain questions, we also conduct statistical data analysis.

Implementation

1. The Keras library is used to implement the final design of our RNN model in Python.

2. For our data, the model uses a fine-tuned Word Embedding Matrix, a ReLu activation function within the hidden layers, and a Sigmoid Function for output.

3. Adam (a Stochastic Gradient Descent extension) is the optimization (training) algorithm used.

4. The Loss Function is the standard Binary Cross Entropy, which simply ensures that our model maximizes the likelihood of our training results.

5. Dropout is used to help prevent our model from being overfitted

6. Instead of a single train-test break, we conduct a 5-fold cross-validation evaluation of our model. In other words, we:-

Implementation

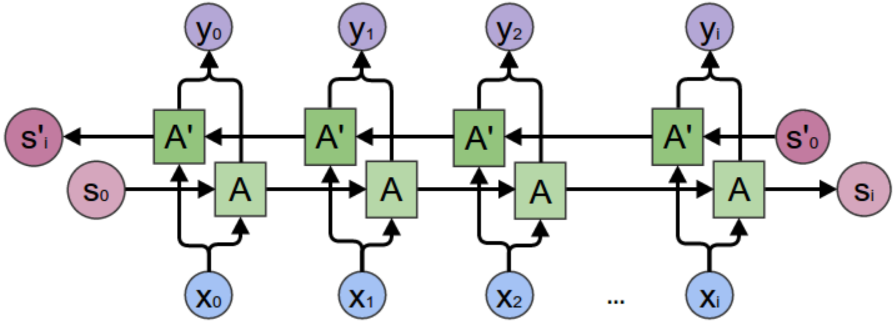
i. Arbitrarily shuffle the dataset.
ii. Divide the data into a total of k classes
iii. For each one-of-a-kind group:
    a. Use the category as a reserve or a test data set.
    b. As a training data collection, use the remaining classes.
    c. Fit a model to the training data and test it on the test data.
    d. Keep the test score but toss out the model.
iv. Finally, using the sample of model evaluation ratings, summarize the model's abilities.

Data Modeling: Recurrent Neural Network/LSTM

- We created a bi-directional Recurrent Neural Network model with LSTM (Long Short Term Memory) modules and Attention Layers, based on the latest research in Natural Language Processing and Deep Learning.
- We wanted to try to replicate the acclaimed success in the Quora Challenge with this dynamic model, which has achieved state-of-the-art results in other NLP tasks.
- The implemented model's bi-directional nature is suitable for this task because we want our RNN to learn representations from previous time steps as well as future time steps in the learning procedure.
- Consider the following question: “Why is the Italian football team so bad?”vs. "Why is Rome the capital of Italy?" Since it is attempting to make a point, the first question will almost definitely be flagged as "Insincere" by Quora's guidelines. On the other hand, the second question is definitely correct. Given that each of these questions begin with the text “Why is the Italian...”, a typical RNN, even with an LSTM implementation, would have difficulty accurately identifying them.

Data Modeling: Recurrent Neural Network/LSTM

- To put it another way, a standard RNN/LSTM will ignore future word sequences, which encapsulate crucial semantic data for the Quora Challenge. A bi-directional network, on the other hand, will relay back instances of words seen further down the line, resolving the issue



Model Evaluation: Recurrent Neural Network/LSTM Results

1. As can be shown, our model achieves an F1 Score of 0.67 across the four Epochs for which it was prepared, with a 96 percent accuracy.

2. We can safely say that the added difficulty, as compared to other modeling methods (such as logistic regression and the nave approach), was well worth it. Since the model achieves a significantly higher F1 score while only slightly increasing overall accuracy.

3. Given that Quora isn't interested in inferring or interpreting what constitutes a "sincere" or "insincere" query, and that the company determines this a priori, we can suggest our "black box" approach as an excellent tool for improving Quora's platform and overall business.

Future Work

There are a number of things we'd like to try if we were to pursue this project in the future. On the RNN side of things, we could extract a lot more features to try to enhance our results. Finally, with more time and computational resources (more efficient GPUs), we might see what happens with BERT with more data and training epochs, which would be interesting considering that BERT showed promising results even with limited data and time.

In this project I did not work on hyperparameter tuning. We can try to improve the performance by performing hyperparameter tuning using hyper opt or Grid Search or Random Search.

Conclusion

For this project our goal was to classify Quora questions as sincere or insincere. We experimented with a number of classification models and optimizations, with mostly positive results. Lstm with Adam optimization and updated decision boundary, as well as regularized Logistic Regression with an adjusted decision boundary, were some of the models we used.

References

1. C. Liu, Y. Sheng, Z. Wei, and Y. Yang, “Research of text classification based on improved tf-idf algorithm,” in 2018IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), Aug 2018, pp. 218–222
2. O. Aborisade and M. Anwar, “Classification for authorship of tweets by comparing logistic regression and naive bayes classi-[U+FB01]ers,” in 2018 IEEE International Conference on Information Reuse and Integration (IRI), July 2018, pp. 269–276.
3. S. T. Indra, L. Wikarsa, and R. Turang, “Using logistic regression method to classify tweets into the selected topics,” in 2016International Conference on Advanced Computer Science and Information Systems (ICACSIS), Oct 2016, pp. 385–390.
4. S.-J. Yen, Y.-S. Lee, J.-C. Ying, and Y.-C. Wu, “A logistic regression-based smoothing method for chinese text categorization.” Expert Systems With Applications, vol. 38, pp. 11 581– 11 590, 2011.
5. A. Onan, S. Koruko˘glu, and H. Bulut, “Ensemble of keyword extraction methods 8and classifiers in text classification.” Expert Systems With Applications, vol. 57, pp. 232 – 247, 2016.
6. “A comparison of several ensemble methods for text categorization,” in IEEE International Conference on Services Computing, 2004. (SCC 2004). Proceedings. 2004, Sep. 2004,pp. 419–422.

Thank You