Shinnosuke Taniya

Pasadena, CA

staniya@g.hmc.edu linkedin.com/in/shinnosuke-taniya github.com/staniya

Project Description

This repository contains deep learning software that I developed with my peer Cindy Lay during the month of November 2021 to perform Stack Overflow question quality classification. The software served to satisfy our final project requirement for a Natural Language Processing (NLP) course at Harvey Mudd College that asked students to develop NLP software and write a research paper presenting their work. Our software automates the quality-based classification of Stack Overflow questions based on their linguistic characteristics and the tags associated with each post.

Kaggle provided the dataset, which contains 60,000 Stack Overflow questions from 2016 to 2020, collected directly from the Stack Overflow website. Our study involved two parts: the first was to apply different text classification techniques: Bi-directional Encoder Representation from Transformers (BERT), Bi-directional Long-Short Term Memory (BLSTM), and Convolutional Neural Networks (CNN) to study and compare their performances using solely the raw body text of posts to predict question quality. The second was to continue off of work done by Bazelli et al. that incorporates sentiment analysis on the raw text of the data to represent text data numerically.[1] Then, Neural Net classification and Random Forest classification were applied, where Random Forest was further used to rank the three provided features of posts: Title, Body, Tag in terms of their feature importance. The three main files in the repository are:

1. Text Classification Using Bert
2. Text Classification Using BLSTM and CNN
3. Sentiment Analysis Incorporated Text Classification Using Neural Net and Random Forest.

The reason for my choice in software is due to my belief that it effectively reveals my understanding of deep learning techniques. With the proliferation of deep learning techniques and their rising importance in AI applications, I concluded that I can best demonstrate my skills by using this project to showcase my data science skills (data structures and data modeling, quantitative analysis methods, statistics, etc.) while also demonstrating my strong foundation of computer science fundamentals (implementing efficient algorithms). Furthermore, the project is coupled with a detailed research paper that discusses our experimental methods, results, and the significance of our findings. Given that our software is written on Google Colaboratory, it is relatively easy to reproduce our findings and follow our experimental process. Lastly, it is important to clarify that although I have worked on more complex software projects such as developing a full-stack hardware cold wallet for EOS, developing an Amazon Alexa service for Cube Wealth, and applying Toshiba's Simulated Bifurcation Machine to find unique solutions to drug discovery problems, I chose this software to avoid violating non-disclosure agreements and because I am confident that it best demonstrates my abilities as a software engineer.

Note: Underlined text are hyperlinks but if they are inaccessible the repository and paper links are:

1. Repository: https://github.com/cindylay/cs159-final-proj
2. Paper: https://github.com/cindylay/cs159-final-proj/blob/main/Stack%20Overflow%20Question%20Quality%20Classification.pdf

[1] B. Bazelli, A. Hindle and E. Stroulia, "On the Personality Traits of StackOverflow Users," 2013 IEEE International Conference on Software Maintenance, 2013, pp. 460-463, doi: 10.1109/ICSM.2013.72.