Application of Deep Learning Frameworks for Classification of Cancer-Related Discussion Posts

Sarah Logan, Thomas Durkin

Motivation

This project will use discussion posts collected from the American Cancer Society's Cancer Survivors Network [1], which is a discussion board for cancer survivors, caregivers, and their families. There are 27 different discussion boards for different cancer types (ex: pancreatic, kidney, etc.) and the goal of this project will be to classify the posts by cancer type using multiple deep learning techniques. There are several motivating factors to this project. First, when a user wants to start a new post, they are required to choose from a long list of discussion boards to post to. With a classification algorithm in place, we could recommend which discussion board the user may want to post to based on the text that they have written, making for a more streamlined user experience. Second, being able to classify the posts will help us to better understand the topics that people who have experienced different cancer types are discussing. Third, we will be using multiple deep learning models and comparing their performance in order to determine which is best able to classify the posts.

Data

The discussion posts from the American Cancer Society's Cancer Survivors Network have already been collected by the team. Only posts that were on a thread with a reply after December 2017 and up until November 2021 were gathered. A total of 102,388 posts have been collected. We will investigate the possibility of collecting posts that have been published in 2022. The existing data will need to be cleaned and vectorized in a way that is conducive for use in the deep learning models.

Modeling Approaches

To accurately classify the discussion posts, we will utilize multiple deep learning models: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Bi-directional Long Short Term Memory (Bi-LSTM), and a Transformer encoder. Model performance will be evaluated using confusion matrices, accuracy, F1 score, and ROC curve. A similar study has been completed to determine disease conditions based on patient's discharge summary notes [2]. We will use their results as a baseline.

Timeline

11/06	Prepare data (clean and vectorize)
11/13	Build initial models
11/20	Look into expanding data
11/27	Re-train models
12/4	Fine tune model performance and compare results

References

[1] American Cancer Society. Cancer Survivors Network. Retrieved October 30, 2022, from https://csn.cancer.org/

[2] Lu, H., Ehwerhemuepha, L., & Rakovski, C. (2022, July 2). *A comparative study on Deep Learning models for text classification of unstructured medical notes with various levels of class imbalance - BMC Medical Research methodology*. BioMed Central. Retrieved October 30, 2022, from https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01665-y