

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

Thomas Durkin & Sarah Logan
December 6, 2022



Motivation

- Cancer is the second leading cause of death in the United States
- Cancer patients seek out support through online discussion forums, such as the American Cancer Society's **Cancer Survivors Network**

New Topic

⊕ Select a discussion board...

Topic Title


Type your message



Cancel

[Save Draft](#)

Post Topic


Cancer Survivors Network

[Discussion Boards](#)
[CSN Chatroom](#)
[cancer.org](#)
[Contact CSN](#)

[Sign In](#)
[Register](#)

- There are 27 discussion boards
- Goal: predict and recommend which discussion board to post to as a user writes their post

Research Questions

- Can we utilize discussion posts from CSN and deep learning frameworks to develop a classifier that predicts which discussion board a post should go to with high accuracy?
- How does our work and our models compare to other research [2] that has been done on classifying text using deep learning?

Data Collection

- Discussion posts were scraped from the Cancer Survivors Network if the thread had a reply between 2018 and 2021
- In total, we collected 102,546 posts
- Discussion boards with $< 1,000$ posts were removed from consideration in our classifier
 - 14 categories were kept

Data Cleaning

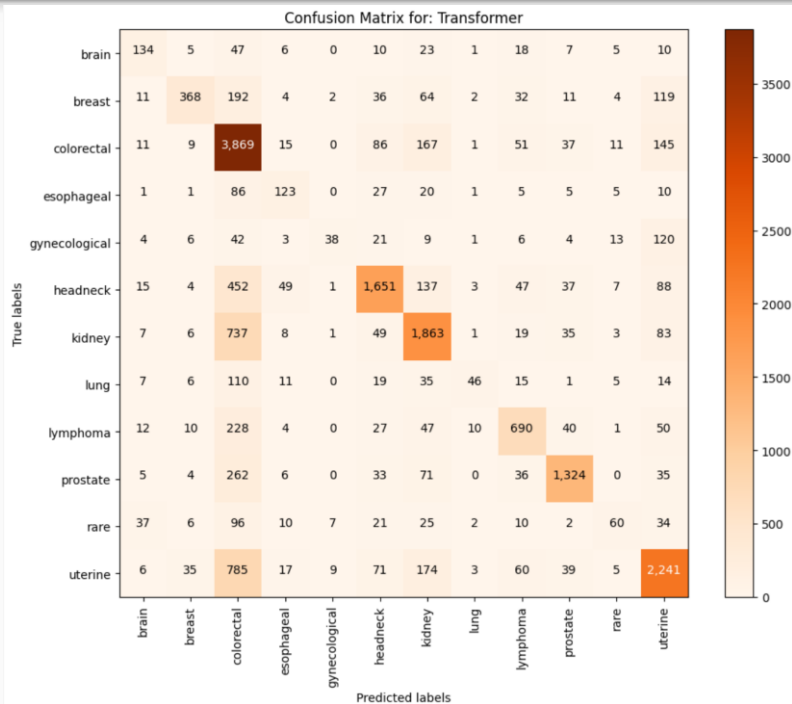
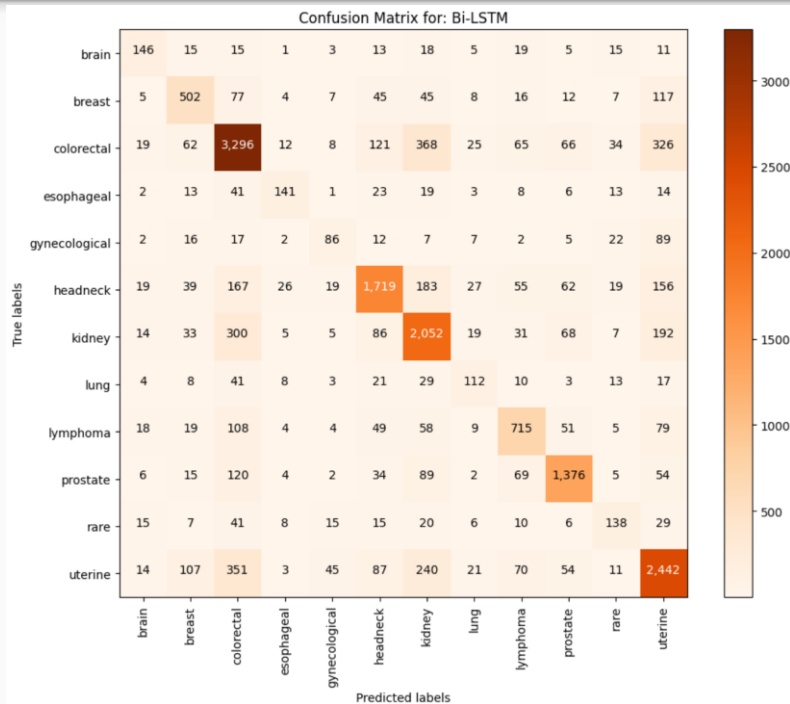
- Emojis were convert to text
- All punctuation was removed
- Remove stopwords and numbers
- Discard all words less than 2 characters

Modeling

- Vocab size of 80,000
- Test size of 20%

Model	Accuracy
CNN	63.8%
RNN	42.7%
Bidirectional LSTM	69.6%
Transformer	67.8%

Results of Top 2 Models



Next Steps

- What is causing the colorectal and uterine categories to falsely predicted?
- Utilize embedding models (GLOVE and SBERT)
- Stacking models

References

1. R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022", *CA: A Cancer Journal for Clinicians*, 2022.
2. H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on Deep Learning models for text classification of unstructured medical notes with various levels of class imbalance", *BMC Medical Research methodology*, 2022.