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

Sarah Logan
University of Rochester
slogan6@ur.rochester.edu

Thomas Durkin
University of Rochester
mdurkin7@ur.rochester.edu

December 20, 2022

## Abstract

In this study, we aimed to construct a deep learning classification model to recommend which discussion board a post should go to after a user has written it on the Cancer Survivors Network, which is a cancer-related public discussion forum. Additionally, we explored multiple types of models and compared their performance on this natural language processing task. We concluded that a stacked model, which was a combination of the Bidirectional LSTM and the transformer encoder outputs, provided the best results with an accuracy of 70.7%.

## 1 Introduction

Today, cancer is ranked as the second leading cause of death in America. It is estimated that almost 2 million people will be diagnosed with cancer in the United States in 2022 [1]. Receiving a cancer diagnosis can be a very scary and confusing time, leading to many cancer patients and their families wanting to speak with others who have similar experiences. One place where cancer patients seek such support is through online discussion forums, such as the American Cancer Society's Cancer Survivors Network [2]. On this website, there are 27 cancer type specific discussion boards that a user could post to. When a user writes a post, they have to scroll through a long list of potential discussion boards and determine which one is most appropriate for their post. The goal of this project is to create a classifier that can accurately recommend which discussion board the user should post to based on the text that they have written to make for a more streamlined user experience on the Cancer Survivors Network. Additionally, this is intended to be a comparative study of deep learning models, such as CNN, RNN, and LSTM, on a natural language processing task.

In recent years, deep learning has becoming increasingly utilized in natural language processing tasks. Deep learning can be advantageous over traditional machine learning methods, such as logistic regression and random forests, because it has the ability to learn features rather than requiring them to be created by hand [3]. For this particular problem, deep learning is more suitable due to the large number of classes (discussion boards) and because, to the human eye, there are many similarities among the posts across different discussion boards, which would make it challenging to engineer effective new features.

## 2 Related Work

While convolutional neural networks were originally intended for use in computer vision, they have also been used in text classification more recently. Some applications include sentiment analysis of movie reviews, classification of sentences as subjective or objective, and classification of questions into different groups [4].

Recurrent neural networks allow text to be pro-

cessed in sequential order and the most common variant is a Long Term Short Term model. LSTM's have proven training and performance gains over regular RNN's due to their ability to handle long-term dependencies. In one study comparing an RNN to an LSTM, their experiments resulted in the LSTM having 8% better relative perpexity than the RNN [5].

A recent study has been conducted by applying multiple text classification techniques to discharge medical notes in order to to determine the disease prevalence of 16 different diseases [6]. The models used include: a Convolutional Neural Network (CNN), a Transformation encoder, and various sequence neural networks. We utilized these models in our own study in order to compare the performance. The results concluded that the Transformer encoder performed the best in almost all testing cases with the CNN giving comparable results only when the disease prevalence is greater than 50 percent.

In another study, the authors propose a combination of deep learning models to improve classification results. They created the model by using a convolution layer, a LSTM layer with attention, a hidden layer and a final softmax layer. They also compared this model to a regular CNN, RNN, and LSTM. They found that the combination of model layers worked best, followed by the LSTM [7]. This has inspired the work done to combine models in this study and allows us to compare whether our model types have similar performance to these ones.

# 3 Methods

The American Cancer Society's discussion forum, the Cancer Survivors Network, is composed of 27 cancer-type specific discussion boards (ex: brain, lung, kidney, etc.) where registered members can post about their experiences with cancer. Both registered and unregistered members are able to view posts on the discussion boards. Discussion posts were scraped using a web scraper that was built using the Beautiful Soup library in Python. Posts were only collected if they were on a thread that had a reply sometime between 2017 and 2021. No personally identifiable information was scraped along with the posts. In total, 102,388 posts were collected to use in the models. Data cleaning was done on the posts in order to remove emojis, punctuation, words with less than two characters, and numbers.

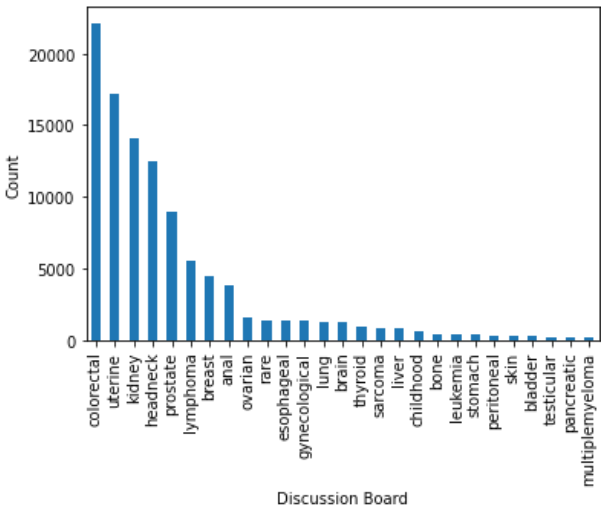


Figure 1: Distribution of posts on each discussion board.

The number of posts collected for each discussion board is shown in Fig. 1. As shown, there is a class imbalance among the number of posts on each of the discussion boards. The colorectal discussion board has the greatest number with 22,137 posts. There are several discussion boards, such as testicular, pancreatic, and multiple myeloma, that have less than 1,000 posts in the data set. Because there are so few training examples for some of these boards, it was decided to only include discussion boards that had over 1,000 posts as classes in the model. This resulted in using a total of 13 classes for prediction.

In order to create an accurate classifier, multiple deep learning models were utilized and evaluated. These included Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Bidirectional Long Short-Term Memory (Bi-LSTM), and a Transformer encoder. Additionally, we tried to combine the Bi-LSTM and the Transformer encoder. The data was split into a training and testing set with a testing set size of 20%. Parameter tuning and tun-

ing of the architecture of the models was done. The overall performance of the models was evaluated on accuracy, F1-score, and class-wise recall.

# 4 Experiments

All models were trained using 8 epochs, a learning rate of 0.001, and a vocab size of 80,000. The loss function used was cross entropy and the optimization algorithm employed was Adam. The final results of the models are shown in Table 1.

| *Model* | $F1 - Score$ | *Accuracy* |
|---|---|---|
| Baseline (decision tree) | 0.5574 | 0.560 |
| CNN | 0.6340 | 0.631 |
| RNN | 0.4008 | 0.395 |
| Bi-LSTM | 0.6912 | 0.687 |
| Transformer | 0.6675 | 0.679 |
| Stacked model | 0.7014 | 0.707 |

Table 1: F1-score and accuracy for each model.

## 4.1 Baseline Model

In order to ensure that using a deep learning model was necessary for this problem, a decision tree model was created as a baseline. This model only had an accuracy of 56.0% and an F1-score of 55.7%, which is significantly worse than almost all of the deep learning models that were tried. This means that the discussion board recommended by this model would only be correct about half of the time and demonstrates that deep learning is a useful tool for this task.

## 4.2 Convolutional Neural Network

The layers used to build the CNN model were an embedding layer, a 1D convolutional layer with a kernel size of two and a stride of one, a max pooling layer, a a fully connected layer, a dropout layer, and a ReLU activation layer. A sequence length of 75 was used. This model resulted in 63.1% accuracy and a 63.4% F1-score. As seen in Table 2, the CNN was not able to accurately identify the actual positive cases for the esophageal, gynecological, lung, ovarian, and rare discussion boards, as it had a recall of 0 for all of those. These boards had some of the fewest training examples in the data set so it makes sense why the model had difficulty predicting these classes. Perhaps if more examples of those classes were collected, the performance could be improved. The CNN did have some success in accurately identifying posts from the colorectal, head and neck, kidney, and prostate discussion boards, as each of these had over a 70% recall. While this performance overall is decent, one reason why convolutional neural network's might not perform as well as other methods on this task is because they aren't able to capture long-term dependencies [3]. On average, the posts in the data set have a 112 words in them and 7 sentences, therefore a CNN may be unable to connect the context across these sentences during training.

## 4.3 Recurrent Neural Network

This model proved to be significantly worse than the other models, only providing an accuracy of 39.5% and an F1-score of 40.0%. Building this model was fairly similar to the CNN, using a RNN layer in place of a convolutional layer and no max pooling layer. The poor performance is due to not only misidentifying class labels but not being able to recognize classes with small amounts of data. The RNN struggled to accurately classify examples from the brain, gynecological, lung, ovarian, and rare discussion boards, with each of these having a recall of 0 (Table 2). Furthermore, it did not do well with any of the other classes, as the recall for each class was never above 66%. Similar to convolutional neural networks, recurrent neural networks can struggle with long-term dependencies, which may explain the poor performance. This is a problem that has long been seen in the practice of training RNN's using gradient descent [8].

## 4.4 Bidirectional Long Short-Term Memory

The Bi-LSTM model had the best performance for an individual model with an accuracy of 68.7% and an

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 0.32 | 0.39 | 0.64 | 0.37 | 0.24 | 0.58 | 0.59 | 0.26 | 0.52 | 0.23 | 0.68 | 0.18 | 0.56 |
| CNN | 0.40 | 0.43 | 0.73 | 0 | 0 | 0.70 | 0.72 | 0 | 0.56 | 0 | 0.72 | 0 | 0.68 |
| RNN | 0 | 0.10 | 0.50 | 0.11 | 0 | 0.28 | 0.66 | 0 | 0.13 | 0 | 0.32 | 0 | 0.44 |
| Bi-LSTM | 0.50 | 0.55 | 0.75 | 0.57 | 0.34 | 0.71 | 0.72 | 0.36 | 0.61 | 0.38 | 0.78 | 0.41 | 0.71 |
| Transformer | 0.34 | 0.41 | 0.83 | 0.33 | 0.15 | 0.71 | 0.72 | 0.30 | 0.64 | 0.10 | 0.77 | 0.17 | 0.69 |
| Stacked model | 0.35 | 0.62 | 0.81 | 0.37 | 0.31 | 0.72 | 0.74 | 0.37 | 0.64 | 0.38 | 0.79 | 0.39 | 0.71 |

Table 2: Recall for each model by class. Classes are shown on the top and are represented by the following numbers - 1: Brain 2: Breast, 3: Colorectal, 4: Esophageal, 5: Gynecological, 6: Head and Neck, 7: Kidney, 8: Lung, 9: Lymphoma, 10: Ovarian, 11: Prostate, 12: Rare, 13: Uterine.

F1-score of 69.1%. The model was constructed with an embedding layer, two Bidirectional LSTM layers, a sigmoid activation layer, two fully connected hidden layers (with a size of 250 and 125 input dimensions, respectively), and a ReLU activation layer. The Bi-LSTM may have performed better than other models because it is a variant of RNN that is specifically tailored to sequential input and it can handle long term dependencies [9]. In particular, it performed well at accurately identifying posts from the colorectal, head and neck, kidney, prostate, and uterine discussion boards (Table 2). Similar to the CNN and RNN, the discussion board that it struggled with the most was the gynecological board, with a recall of only 34%.

### 4.5 Transformer Encoder

The transformer encoder model had decent accuracy at 67.9% and an F1-score of 66.8%. This model was constructed with an embedding layer, a positional encoder layer, a transformer encoder layer with 10 attention heads, and a fully connected layer. The transformer encoder just slightly underperformed the Bi-LSTM on recall in all classes except colorectal and lymphoma (Table 2). In the classes where the transformer encoder outperformed the Bi-LSTM, it had a higher recall by an average of 5.5 percentage points. In the study previously discussed in [6], they found that the transformer produced the best results, however, for our problem, the transformer did well but was not the best model on its own. To further improve the two best models, we decided to combine them and the results are discussed below.

### 4.6 Stacked Model

The best results were achieved by stacking the outputs of the two best models, the Bidirectional LSTM and the transformer encoder. With this technique we were able to attain an accuracy of 70.7% and an F1-score of 70.1%. The stacked model had the same or better recall in comparison to the best LSTM model in all classes except brain, esophageal, gynecological, and rare. Thus, it can be concluded that stacking the models definitely makes improvements on the predictions over just the LSTM or just the transformer encoder. While we didn't combine our models in the same way as the authors in [7], we did find that the combination of an LSTM and an attention mechanism performed the best. Additionally, our finding that the LSTM was the next best performing model on the text classification task after the combined model was consistent with that study.

## 5 Conclusion

In this study, multiple deep learning models were used to classify which discussion board a post belongs to on the American Cancer Society's discussion forum, the Cancer Survivors Network. The performance of multiple models, including CNN, RNN, LSTM, and a transformer encoder, was compared. Additionally, after observing that the best performing models were the LSTM and the transformer en-

coder, we attempted to combine these models. Ultimately, the two models that performed best were the Bidirectional LSTM and the stacked model. It is reasonable to believe that these models performed better than other deep learning models, such as CNN and RNN, because they are able to handle long-term dependencies across lengthy posts. All of the models did well at accurately predicting posts from the colorectal, kidney, and uterine discussion boards, but they all also faced difficulty in predicting posts from the gynecological, lung, ovarian, and rare discussion boards. These boards each have less than 1550 posts so it may be worthwhile to collect additional posts from the boards in the future to improve the classification models. Overall, implementing the stacked model on the Cancer Survivors Network to recommend which discussion board a user should post to after they have written their post would make for a better user experience and allow users to more easily get the support they need.

# References


[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal. Cancer statistics, 2022, *CA: A Cancer Journal for Clinicians*, 2022. 1

[2] Cancer Survivors Network. Available: https://csn.cancer.org/. [Accessed: 30-Oct-2022]. 1

[3] T. Young, D. Hazarika, S. Poria and E. Cambria. Recent Trends in Deep Learning Based Natural Language Processing [Review Article], *IEEE Computational Intelligence Magazine*, 13 (3): 55-75, 2018, doi: 10.1109/MCI.2018.2840738. 1, 3

[4] K. Yoon. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014. 1

[5] M. Sundermeyer, R. Schluter, and H. Ney. Lstm neural networks for language modeling, *Interspeech*, pages 194–197, 2012. 2

[6] 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. 2, 4

[7] X. Bai. Text classification based on LSTM and attention. In *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, pages 29-32, 2018, doi: 10.1109/ICDIM.2018.8847061. 2, 4

[8] Y. Bengio, P. Simard and P. Frasconi. Learning long-term dependencies with gradient descent is difficult, *IEEE Transactions on Neural Networks*, 5 (2): 157-166, 1994, doi: 10.1109/72.279181. 3

[9] S. Hochreiter, J. Schmidhuber. Long short-term memory, *Neural Compututation*, 9 (8): 1735-1780, 1997. 4