Machine Learning Approaches to Classify Topics in Forums

Sitong Liu
stliu@unc.edu

1 Introduction

Against the background of today's rapid Internet development, online forums have become an important platform for university students to discuss course content and communicate about campus life. Triple Uni is a typical example, a forum shared by students from the University of Hong Kong (HKU), the Chinese University of Hong Kong (CUHK) and the Hong Kong University of Science and Technology (HKUST). As the forum expands and the number of posts increases, content retrieval is challenged by the massive amount of information.

One solution is to encourage users to actively tag posts with topics so that other users can more accurately find the information they need. However, the reality is that not all users are happy to actively tag posts. Many users, when asked by the system, just randomly select a topic from the list of topics to pass the validation, which reduces the differentiation and usefulness of topic tags.

In this study, machine learning approaches are used to automatically classify posts with the aim of improving classification accuracy and optimizing user experience. Specific methods include training three classifiers, random forest classifier, softmax classifier and neural network classifier, and comparing their performance to determine the most suitable machine learning model for this forum.

2 Dataset and Data Cleaning

2.1 Dataset Description

The dataset collected for this study is from the Triple Uni and includes half a million posts numbered 1 to 500000 and user-selected topics on the platform. This study applies the text-embedding-ada-002 model provided by OpenAI to embed the post text and generates a 1536-dimensional representation vector. To follow the privacy rules of the platform, we desensitize the raw text in the dataset and keep only the corresponding text-embedding vectors. The structure of the dataset is as follows:

Field Name	Type	Description
uni_post_id	int64	Post id of the forum
post_topic	string	Topic of the post (in Chinese)
embedding	float[1536]	Embedding of the post content

Table 1: Dataset Structure

2.2 Data Cleaning

In order to cope with the change in the platform's topic categorization, we decide to select the target topics as Trading, Academics, Emotions, Job Hunting and Random Thoughts. In this study, we re-classify the topics with similar meanings to these target topics into the designated topics and exclude all other topics. Given that the original topics are in Chinese, we also convert these topics into English representations. The specific topics are classified as follows:

Topic in English	Count
Trading	15963
Academics	88468
Emotions	44848
Job Hunting	11105
Random Thoughts	141586

Table 2: Topics List

2.3 Data Splitting

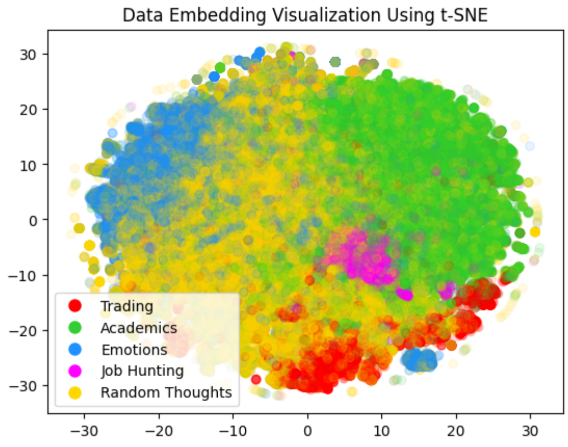
The dataset is divided into two parts: 80% for the training set and 20% for the test set. The aim of this move is to verify that our model is able to show excellent performance on unseen data, rather than just overfitting on the training data.

2.4 Data Accessibility

The dataset for this study is publicly available at https://drive.google.com/drive/u/1/folders/1VVpugGyS-9-mhkvHh4wTX-feFAt3hEVP.

3 A Closer Look at the Data

In order to visualize the relationship between embedding and topic categories, this study adopts the t-SNE method to reduce embedding from high dimensionality to 2 dimensionality while preserving the relative position information. By using different colors to identify each topic, we achieve a visual presentation of the data.



We observe that posts on the same topics tend to cluster together in the graph, which validates the feasibility of classifying posts using machine learning's embedding technique. In particular, posts on the topics of Job Hunting and Academics form tight clusters in the graph, showing a high degree of correlation between posts of these topics, signaling that these topics may perform better in the classification task. On the contrary, posts on the topic of Random Thought are more scattered in the graph, which may be due to the fact that Random Thought covers a wider range of content, and other posts that are not easy to categorize may be classified as this topic, making it more difficult to categorize them.

Since topics are specified by users, their accuracy cannot be guaranteed. As a result, many outliers appear on the graph, and these may indicate errors in the topic labeling of posts. How to minimize the impact of these mislabeled data is a major challenge for us. In addition, the uneven number of posts across topics poses an additional difficulty to the classification task.

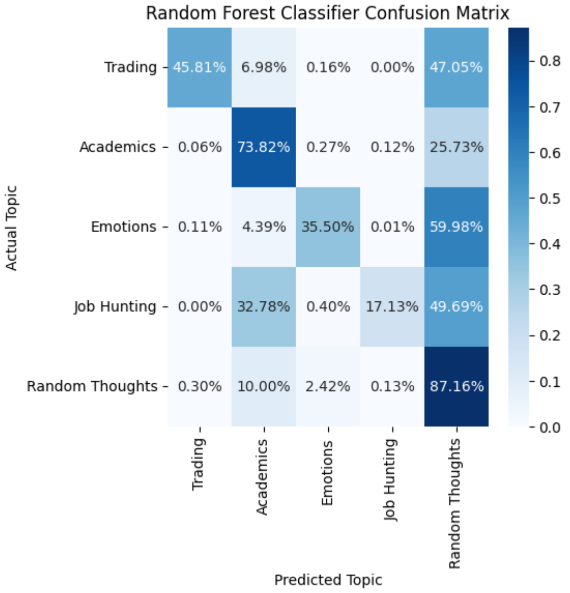
4 Models and Training

In order to select the best model, this study analyzes the data using random forest, softmax regression and neural network respectively and compares their accuracy.

4.1 Random Forest

According to Breiman (2001), random forest models are very effective in dealing with high dimensional data, can effectively prevent overfitting and are more resistant to interference with mislabeled training data.

We train a random forests model with 100 estimators using the training set and test the model on the test set. On the test set, the model achieved an accuracy of 70.9027%.



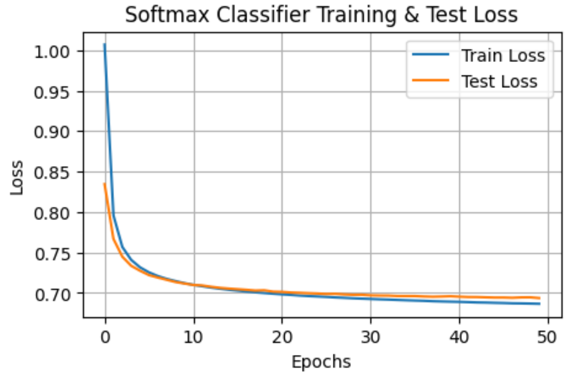
Although the accuracy is not low, we find that random forest model does not handle the unbalanced training data well. Nearly half of the training data is in the Random Thought category, which leads to a tendency for the model to predict most inputs as Random Thought topics, and thus random forest does not perform well on this dataset.

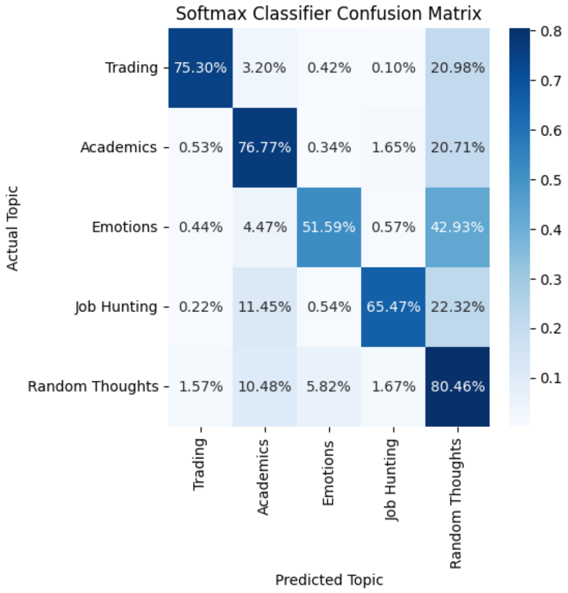
4.2 Softmax Regression

Bishop (2006) describes softmax regression as a generalization of logistic regression that is applicable to multi-class classification problems, where it calculates the probabilities of each class over all possible classes.

In order to solve the problem of unbalanced training data, we optimize the cross-entropy loss function and introduce a weighted loss function based on the proportion of categories. This function takes into account the proportions of different categories and imposes higher penalties for classification errors in niche categories, thus improving the model's classification of all categories.

We set the batch size to 256, the learning rate to 0.001 and set the epochs to 50 and use these parameters to train the softmax model. The model achieves 74.2921% accuracy on the test set.



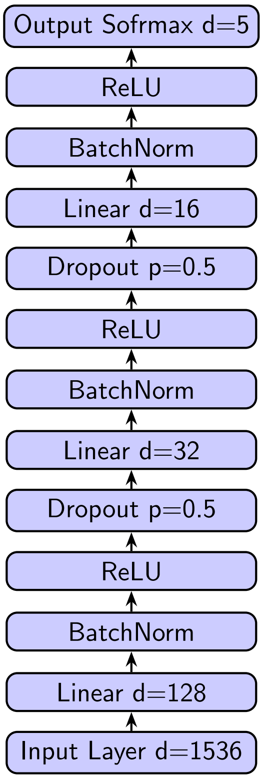


The softmax model achieves higher accuracy compared to random forest and effectively mitigates the unbalanced training data problem. Moreover, due to the simpler model structure, softmax is significantly faster than random forest in both training and inference. All these indicate that softmax is the superior choice in this problem.

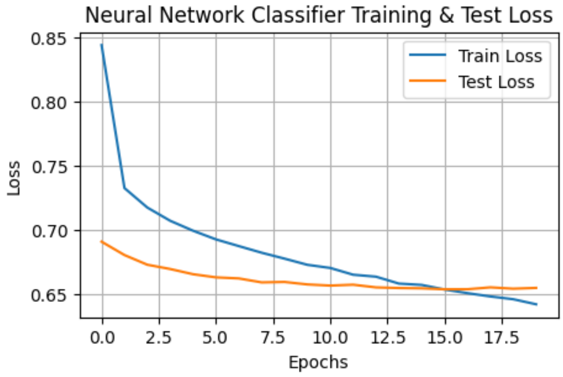
4.3 Neural Network

As highlighted by LeCun, Bengio, and Hinton (2015), neural networks have been pivotal in advancing classification tasks across various domains, harnessing their ability to learn complex patterns through deep architectures.

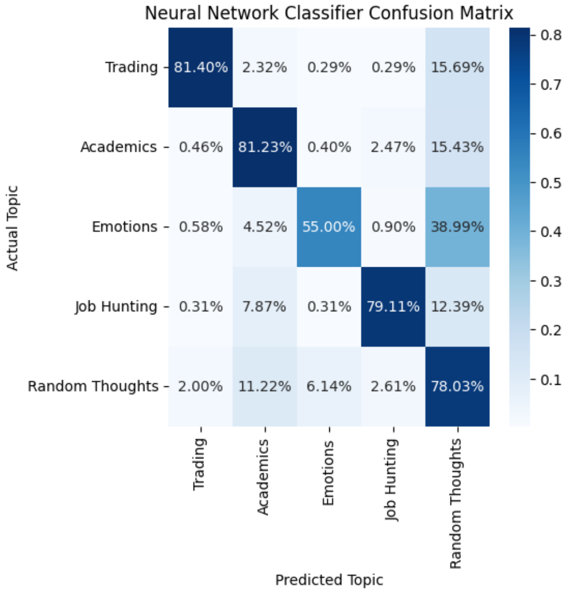
We use the most basic neural network model, which consists of three fully connected layers, each followed by a batch normalization and ReLU activation function. To prevent overfitting, a Dropout layer with a dropout rate of 0.5 is added after the first and second fully connected layers. The structure of the model starts with an input layer, first a fully connected layer that maps the input dimensions to 128, then through a 32-dimensional and a 16-dimensional fully connected layer, and finally through an softmax layer that produces predictions.



We train the neural network model with a batch size of 256, a learning rate of 0.001, and with 20 training cycles. With these settings, the model achieves 75.7923% accuracy on the test set.



There is a further improvement in the model accuracy of the neural network, while the unbalanced training data problem is further mitigated. This suggests that increasing the depth of the network helps to improve the model's ability to gener-



alize. However, the model is still more difficult to distinguish between Random Thought and Emotions, which may be due to the fact that the content of the posts under these two topics is relatively similar without clear boundaries. In addition, the training and inference speed of the neural network model is significantly slower than that of the softmax model, and in practical applications, we need to consider the cost and effect together to choose the most suitable model.

5 Summary

5.1 Models Comparison

Model	Test Accuracy
Random Forest	70.9027%
Softmax Regression	74.2921%
Neural Network	75.7923%

Table 3: Models Comparison

5.2 Conclusion

In this study, we test the performance of multiple machine learning techniques on a topic classification task. The results show that the model containing a softmax layer outperforms the random forest model, while the neural network model further outperforms the softmax regression model in terms of accuracy. Nevertheless, the overhead of neural networks is significantly higher than that of the softmax regression model, both in terms of training and inference, making the latter a more cost-effective option.

Further improvement of model accuracy is challenged by the limitations of data labeling. One potential solution is to manually calibrate the data labels of the test set, which would provide more reliable metrics for evaluating model effectiveness. However, due to time and resource constraints, this method could not be implemented in this study, which would be an important direction for future research.

5.3 Live Demo

We develop an live demo application using Streamlit which allows the user to enter a text and shows the classification results of different models. You can experience the performance of these models by visiting https://tree-hole-judge.tripleuni.com/.

5.4 Code Accessibility

The code for our data cleaning, model training, as well as the live demo, are all open source and available at the following GitHub link https://github.com/lststar/Triple-Uni-Topic-Classifier.

References

[1] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324.

[2] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* https://doi.org/10.1007/978-0-387-45528-0.

[3] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. https://doi.org/10.1038/nature14539.