Final Project Proposal

Chinese News Classification using BERT:

1. Introduction:

The aim of this project is to develop a text classification model for Chinese language news articles using the BERT (Bidirectional Encoder Representations from Transformers) model. The model will be trained to classify news articles into predefined categories, enabling efficient and automated news analysis.

2. Background:

Text classification has become an essential tool in the field of journalism and media, where it can be used for automated news categorization, sentiment analysis, and topic labeling. BERT, a transformer-based machine learning technique for natural language processing (NLP), has been proven to achieve state-of-the-art results in a variety of NLP tasks.

3. Methodology:

The project will be carried out in the following steps:

1. **Data Collection**: Collect a large dataset of Chinese news articles along with their respective categories. The data should be diverse and representative of the different types of news categories.
2. **Preprocessing**: Preprocess the data by cleaning, normalizing, and tokenizing the texts. The Chinese texts will be segmented into words.
3. **Model Training**: Fine-tune a pre-trained BERT model on the collected dataset. The model will learn to understand the context of Chinese words and predict the category of a news article.
4. **Evaluation**: Evaluate the model’s performance using metrics such as accuracy, precision, recall, and F1-score.

4. Expected Outcomes

The expected outcome of this project is a robust Chinese text classification model that can accurately classify news articles into their respective categories. This will greatly enhance the efficiency of news analysis and provide valuable insights into the distribution and trends of news topics.