Turkish News Classification Project Report

Introduction

This report details a comprehensive project on Turkish news classification using state-of-the-art natural language processing models. The project aims to classify Turkish news articles into various categories using three different models: BERT, ELECTRA, and GPT-2. The dataset used for this project is the TR-News dataset, which contains a large collection of Turkish news articles with their corresponding categories.

Objectives

The specific objectives of the project include:

- To accurately classify Turkish news articles into relevant categories.
- To analyze the performance of different natural language processing models.
- To identify challenging categories and develop strategies for improved classification accuracy.

Dataset Preparation

The TR-News dataset was loaded and preprocessed to ensure optimal performance for the classification task. The preprocessing steps included:

1. Cleaning and grouping categories to reduce the number of classes and ensure a sufficient number of samples per class.
2. Selecting the top categories based on the number of samples.
3. Removing certain categories ("Gündem" and "Yaşam") to focus on more specific news topics.
4. Limiting the dataset size to 50,000 samples to manage computational resources effectively.

After preprocessing, the final dataset contained the following categories:

1. Türkiye
2. Dünya
3. Spor

4. Ekonomi
5. Sağlık
6. Kültür-Sanat
7. Eğitim
8. Teknoloji

Model Architecture and Training

Three different models were used for this classification task:

1. **BERT** (dbmdz/bert-base-turkish-cased)
2. **ELECTRA** (dbmdz/electra-base-turkish-cased-discriminator)
3. **GPT-2** (ytu-ce-cosmos/turkish-gpt2-large)

All models were fine-tuned on the preprocessed dataset using the Hugging Face Transformers library. The training process included:

- Using a custom `ContiguousTrainer` class to ensure proper memory management.
- Implementing early stopping and model checkpointing to save the best-performing model.
- Utilizing mixed-precision training (FP16) to improve training speed and efficiency.
- Employing gradient accumulation to simulate larger batch sizes.

Training hyperparameters:

- Number of epochs: 3
- Learning rate: 2e-5
- Weight decay: 0.01
- Warmup steps: 500
- Batch size: 32 for BERT and ELECTRA, 16 for GPT-2 (due to its larger size)

Results and Evaluation

The models were evaluated on a held-out validation set, and their performance was measured using accuracy and a detailed classification report. Here are the results for each model:

BERT

Accuracy: 0.8827

Classification Report:

```
             precision    recall  f1-score   support

       Dünya       0.89      0.91      0.90      1749
     Ekonomi       0.85      0.83      0.84       855
      Eğitim       0.83      0.83      0.83       235
Kültür-Sanat       0.74      0.69      0.71       402
      Sağlık       0.87      0.89      0.88       637
        Spor       0.96      0.98      0.97      1332
   Teknoloji       0.67      0.67      0.67       323
     Türkiye       0.89      0.89      0.89      4467

    accuracy                           0.88     10000
   macro avg       0.84      0.83      0.84     10000
weighted avg       0.88      0.88      0.88     10000
```

ELECTRA

Accuracy: 0.8725

Classification Report:

```
             precision    recall  f1-score   support

       Dünya       0.88      0.90      0.89      1749
     Ekonomi       0.85      0.81      0.83       855
      Eğitim       0.82      0.83      0.83       235
Kültür-Sanat       0.71      0.70      0.70       402
      Sağlık       0.87      0.89      0.88       637
        Spor       0.96      0.97      0.97      1332
   Teknoloji       0.68      0.54      0.60       323
     Türkiye       0.87      0.88      0.88      4467

    accuracy                           0.87     10000
   macro avg       0.83      0.82      0.82     10000
weighted avg       0.87      0.87      0.87     10000
```

GPT-2

Accuracy: 0.8883

Classification Report:

```
             precision    recall  f1-score   support

       Dünya       0.89      0.92      0.90      1749
     Ekonomi       0.84      0.81      0.82       855
      Eğitim       0.86      0.74      0.79       235
Kültür-Sanat       0.77      0.74      0.76       402
      Sağlık       0.87      0.88      0.87       637
        Spor       0.97      0.98      0.97      1332
   Teknoloji       0.75      0.72      0.73       323
     Türkiye       0.90      0.90      0.90      4467

    accuracy                           0.89     10000
   macro avg       0.85      0.84      0.84     10000
weighted avg       0.89      0.89      0.89     10000
```

Discussion

All three models performed well on the Turkish news classification task, with accuracies ranging from 87.25% to 88.83%. The GPT-2 model achieved the highest overall accuracy of 88.83%, followed closely by BERT at 88.27% and ELECTRA at 87.25%.

Key observations:

1. **Performance by Category:** All models performed exceptionally well on the "Spor" (Sports) category, with F1-scores of 0.97. This suggests that sports-related news articles have distinct features that make them easily distinguishable from other categories.
2. **Challenging Categories:** The "Teknoloji" (Technology) category proved to be the most challenging for all models, with the lowest F1-scores. This could be due to potential overlap with other categories or a lack of distinct linguistic features in technology-related articles.
3. **Model Comparison:** GPT-2 showed slight improvements in precision and recall for most categories compared to BERT and ELECTRA, which may be attributed to its larger model size and more extensive pre-training on Turkish text.

4. **BERT vs. ELECTRA:** BERT and ELECTRA performed similarly, with BERT having a slight edge in overall accuracy. This suggests that the additional pre-training task in ELECTRA (replaced token detection) did not provide a significant advantage for this particular classification task.
5. **Cultural Categories:** All models struggled somewhat with the "Kültür-Sanat" (Culture-Art) category, possibly due to its diverse content and potential overlap with other categories.

Conclusion

This project demonstrates the effectiveness of large pre-trained language models for Turkish news classification. The GPT-2 model showed the best overall performance, but all three models achieved high accuracies, indicating their suitability for this task.

Future work could focus on:

1. **Ensemble Methods:** Experimenting with ensemble methods to combine the strengths of different models.
2. **Hyperparameter Tuning:** Fine-tuning hyperparameters to potentially improve performance further.
3. **Challenging Categories:** Investigating the challenging categories (e.g., Technology and Culture-Art) to understand the reasons for lower performance and develop strategies to improve classification accuracy.
4. **Recent Models:** Exploring the use of more recent models like GPT-3 or T5 for Turkish text classification tasks.

Overall, this project provides a strong foundation for Turkish news classification and demonstrates the applicability of transfer learning techniques to Turkish natural language processing tasks.